وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR UNIVERSITY -ANNABA-
UNIVERSITE BADJI MOKHTAR -ANNABA-

جـامعة بـاجي مختار
–عنـابة–

Faculté : Sciences de l'Ingéniorat

Département : Informatique

-Année 2019-

THESE

Présentée en vue de l'obtention du diplôme de **Doctorat en Sciences**

Analyse et Reconnaissance des Activités Humaines à partir des Séquences Vidéo

Option : Informatique

Par

LADJAILIA Ammar

Directeur de thèse	BOUCHRIKA Imed	Pr.	Université de Souk-Ahras, Algérie
Co-Directeur de thèse	MEROUANI Farida Hayet	Pr.	Université de Annaba, Algérie
	Devant le jury		
Président	SOUICI-MESLATI Labiba	Pr.	Université de Annaba, Algérie
Examinateur	BAHI Halima	Pr	Université de Annaba, Algérie
Examinateur	BOUKROUCHE Abdelhani	Pr	Université de Guelma, Algérie
Examinateur	BEKHOUCH Amara	MCA	Université de Souk-Ahras, Algérie

UNIVERSITÉ BADJI MOKHTAR-ANNABA-
FACULTÉ DES SCIENCES DE L'INGÉNIORAT
DÉPARTEMENT DE L'INFORMATIQUE

Analyse et Reconnaissance des Activités humaine à partir des Séquences Vidéo

Présentée en vue de l'obtention du diplôme de DOCTORAT en **Informatique**, préparé par :

Ammar LADJAILIA

Directeur de thèse : Imed BOUCHRIKA
Co-Directeur de thèse : Farida Hayet MEROUANI

2019

ملخص

المحور الرئيسي لهذه الأطروحة البحثية هو التعرف والتحليل الآلي للأنشطة البشرية من خلال مقاطع الفيديو وهذا من أجل تحديد الأنشطة البشرية التي حدثت من خلال دراسة محتواها. لقد برز هذا الموضوع البحثي في الآونة الأخيرة على أنه موضوع أساسي في مجال رؤية الكمبيوتر وتعلم الآلة .هذه الاشكالية هي صعبة للغاية نظرا لوجود اختلافات كبيرة في المظهر والحركة وكذلك عند أداء الأعمال بالإضافة إلى عوامل أخرى متعلقة بزاوية الحصول على الفيديو والخلفيات المشوشة وكذلك حجب الشيء المتحرك. والامر يزداد تفاقما بسبب الكم الهائل من بيانات الفيديو المطلوب تحليلها .إن العديد من التطبيقات في حاجة ماسة إلى حل قائم على رؤية الكمبيوتر للمساعدة في الفهم الآلي للأفعال البشرية ونذكر من ذلك الأمن والرياضة وحتى من أجل السيارات ذاتية التحكم.

في هذه الرسالة، اقترحنا واصفًا للحركة مستنبطا من حسابنا للتدفق البصري وهذا للتعرف على النشاط البشري، هذا مع مراعاتنا للخصائص المستمدة من الحركة فقط. ترميز النشاط البشري يتألف من رسم بياني يضم على ميزات حركية التي تشمل سمات محلية وأخرى شاملة. من أجل الدراسات المقارنة وكذلك تحليل أداء النظام، تم تطبيق أنواع مختلفة من المصنفات، بما في ذلك KNN وشجرة القرار وكذلك SVM على الواصف المقترح، وفي نفس السياق استعملنا أيضا التعليم المعمق. وأجريت أيضا مزيدا من التحاليل لتقييم الواصف المقترح تخص دقة الصور وإزالة بعض الصور، تتفق النتائج المحرزة مع الدراسات النفسية المبكرة التي تفيد بأن الحركة البشرية كافية لفهم الأنشطة البشرية.

الكلمات المفتاحية: التعرف على الأنشطة البشرية، واصف الحركة، التدفق البصري، تجزئة الأنشطة.

Abstract

The main focus of this research thesis is the automated recognition and analysis of human activities from video sequences in order to determine what human actions occur. This area has recently emerged as a fundamental research topic in the field of computer vision and machine learning. This problem is particularly difficult due to the huge variations in the appearance and motion variations when performing actions in addition to challenging factors related to the acquisition settings as viewpoint, background clutters and occlusions. This is further exacerbated by the huge amount of video data to analyse. Many applications are related to this field such as safety, health and sport, as well as autonomous cars.

In this thesis, we propose a motion descriptor based on optical flux estimation for human action recognition, taking into account only the characteristics derived from motion. The signature of human action consists of a histogram containing kinematic features that include local and global traits. Experimental results from the Weizmann and UCF101 datasets confirmed the potential of the proposed approach with classification rates of 98.76% and 70% respectively to distinguish between different human actions. For comparative and performance analysis, different types of classifiers including KNN, decision tree and SVM are applied to the proposed descriptors and deep Learning was also used. Further analysis is performed to assess the proposed descriptors under different resolutions and frame rates. The obtained results are in alignment with the early psychological studies reporting that human motion is adequate for the perception of human activities.

Keywords : Human Action Recognition, Motion Descriptor, Optical Flow, Decomposing Activities

Résumé

L'objectif principal de cette thèse est la reconnaissance et l'analyse automatiques des activités humaines à partir de séquences vidéo afin de déterminer quelles sont les activités humaines qui se produisent. Ce domaine s'impose récemment comme un thème de recherche fondamentale dans le domaine de la vision par ordinateur et l'apprentissage automatique. Ce problème est particulièrement difficile en raison d'énormes variations dans les aspects visuels et de mouvement des personnes et des actions, les changements de point de vue, le fond mobile, des occlusions, la présence de bruits, ainsi que l'énorme quantité de données vidéo. De nombreuses applications sont liées à ce domaine telles que la sécurité, la santé et le sport, ainsi que les voitures autonomes.

Dans cette thèse, nous proposons un descripteur de mouvement basé sur l'estimation de flux optique pour la reconnaissance des actions humaines en ne prenant en compte que les caractéristiques dérivées du mouvement. La signature de l'action humaine se compose d'un histogramme contenant des caractéristiques cinématiques qui incluent les caractères locaux et globaux. Les résultats expérimentaux réalisés sur les bases de données Weizmann et UCF101 ont confirmé le potentiel de l'approche proposée avec des taux de classification atteints de 98,76% et 70% respectivement pour distinguer les différentes actions humaines. Pour l'analyse comparative et l'analyse de la performance, différents types de classifieurs sont utilisés tels que KNN , Arbre de décision et SVM à la base du descripteur proposé et on a aussi utilisé le Deep Learning. Une analyse plus approfondie est effectuée pour évaluer ce descripteur en fonction des différents types de contraintes telles que la faible résolution de la vidéo et le saut des trames. Les résultats obtenus sont en harmonie avec les premières études psychologiques indiquant que le mouvement humain est adéquat pour la perception des activités humaines.

Mots clés : Reconnaissance des actions humaines, Descripteur de mouvement, flux optique, décomposition des activités.

Remerciements

Au terme de cette thèse, je tiens en premier lieu à exprimer ma profonde gratitude à mon Dieu qui m'a donné la force, la volonté, le courage, et surtout la patience pour terminer ce travail. Nombreuses sont les personnes qui ont contribué à réaliser ce travail, auxquelles je présente avec plaisirs mes sincères remerciements.

Je tiens tout d'abord à exprimer toute ma reconnaissance et mon profond respect à ***Pr. BOUCHRIKA Imed*** pour son encadrement de thèse et pour sa confiance et son soutien. Ses travaux n'auraient pas été possibles sans ses conseils, ses encouragements et les nombreuses relectures à toute heure du jour et de nuit. Et je remercie son soutien et ses précieux conseils qui m'ont permis de mener ce travail dans de très bonnes conditions. Merci beaucoup ***Pr. BOUCHRIKA***.

Je souhaiterais ensuite remercier mon Co-encadreur ***Pr. Farida Hayet MEROUANI*** pour son encadrement de thèse et pour sa confiance, son soutien et encouragements. Merci beaucoup Pr. MEROUANI.

Un grand merci au ***Pr. Mourad Zayed*** et ***Dr. Ridha Ejbali***, superviseurs de mon stage à l'Université de Gabès en Tunisie pour leurs patiences et leurs contributions significatives qui a mené à la réussite de cette formation.

Un grand merci à Mme. ***SOUICI-MESLATI Labiba*** professeur à l'université d'Annaba de m'avoir fait l'honneur de présider le jury de ma thèse.

Je suis également très reconnaissant à Mme. ***BAHI Halima*** professeur à l'université d'Annaba de m'avoir fait l'honneur d'être membre du jury de ma thèse.

Je suis également très reconnaissant à M. ***BOUKROUCHE Abdelhani*** professeur à l'université de Guelma de m'avoir fait l'honneur d'être membre du jury de ma thèse.

Je suis également très reconnaissant à M. ***BEKHOUCH Amara*** maître de conférence à l'université de Souk-Ahras de m'avoir fait l'honneur d'être membre du jury de ma thèse.

Et enfin, je veux remercier tous ceux qui m'ont aidé de près ou de loin dans l'élaboration et la finalisation de ce modeste travail.

Dédicace

Je dédie ce travail à :

- *Mes parents pour l'amour qu'ils m'ont toujours donné*
- *À ma femme BOUZID AIDA,*
- *À mes enfants Asma, Ibrahim, Zakaria, Abdelmouiz, Abdelrraouf,*
- *Tous ceux qui veulent partager ma joie...*

Ammar

Table des matières

Table des figures

Liste des tableaux

1 Introduction générale

1.1 Contextes et Motivations

Cette thèse s'inscrit dans le contexte de l'analyse et la reconnaissance d'activités humaines à partir des séquences vidéo. Récemment, une grande partie de la recherche scientifique en vision par ordinateur est consacrée à l'analyse des mouvements humains à partir des vidéos. Ces études concernent un grand nombre d'applications où l'analyse automatique du mouvement humain est jugée cruciale, notamment en biométrie, la surveillance automatique et intelligente, l'arbitrage sportif, les interactions homme-machine, les robots et les automobiles autonomes [123, 132, 23, 151]. Alors que nous sommes de plus en plus des natifs du numérique à l'ère moderne, la reconnaissance des activités humaines devient un domaine de recherche intéressant, capable de s'intégrer dans divers contextes réalistes centrés sur l'être humain [1, 38]. En outre, en raison de l'augmentation sans précédent des données multimédias générées en continu par les caméras de sécurité, la production de films et les téléchargements sur le Web, il devient désormais indispensable d'analyser ce contenu vidéo de manière sémantique par le biais de méthodes automatiques. Cela constitue une étape importante pour faciliter le processus d'indexation, de la recherche et de la récupération de contenu multimédia. L'utilisation de systèmes de vision automatisés pour reconnaître les activités humaines peut constituer une solution innovante pour accroître l'adoption et la convivialité de telles applications visuelles intelligentes.

1.2 Contraintes techniques

Le processus d'extraction et de reconnaissance des actions humaines via des méthodes automatiques sans marqueur sont deux tâches distinctes qui s'affirment encombrantes et complexes. Plusieurs approches sont proposées dans la littérature, ces approches peuvent être basées sur des équipements spéciaux installés sur la personne, y compris des capteurs [135]. L'inconvénient de ces méthodes est qu'elles sont loin de la réalité et ne peuvent être réalisées que dans des laboratoires. Les autres approches qui sont basées sur les méthodes de la vision par ordinateur proposent des solutions qui sont encore en stade de recherche, ceci est dû principalement au degré élevé de liberté du corps humain, associé à la variabilité imprévisible de l'apparence ; cela aggraverait les défis supplémentaires liés à l'étape d'extraction des caractéristiques [156]. Les difficultés peuvent provenir de l'environnement d'acquisition qui comprend l'éclairage, le bruit de fond, le point de vue et le mouvement de la caméra ainsi que l'auto-occlusion ou l'occlusion d'autres objets. Pour le dernier facteur, les personnes peuvent effectuer la même activité de différentes manières et façons [241]. Cela dépend de la culture, du contexte ou des personnes elles-mêmes. De plus, une activité spécifique réalisée par différentes façons peut avoir une sémantique totalement différente et sans rapport. Plus difficiles encore, la plupart des activités humaines se déroulent en parallèle et s'entrelacent, par exemple, un sujet peut utiliser un ordinateur de bureau tout en mangeant ou en parlant au téléphone en même temps.

1.3 Principales contributions

Le but de ce manuscrit est de proposer une approche de reconnaissance des actions humaines de manière automatique à partir des séquences vidéo et une méthode d'analyse de performance du système proposé. Les principales contributions de cette thèse peuvent être résumées dans les deux points suivants :

A) proposition deux architectures pour la reconnaissance des actions humaines ; la première est basée sur le descripteur de mouvement et la deuxième s'articule sur la représentation 2D des actions et la classification à l'aide de Deep Learning.

1. La première approche est basée sur un descripteur de mouvement. Ce descripteur est basé uniquement sur les caractéristiques des mouvements extraites à partir des séquences vidéo. Ses caractéristiques sont dérivées de l'estimation du flux optique à partir d'un triplé des trames consécutives afin de générer un histogramme de mouvement. Le descripteur proposé est composé d'un ensemble des données cinématiques décrivant les propriétés globales et locales des mouvements tirées par l'estimation de flux optique. Ce qui distingue notre descripteur est que la phase d'extraction des caractéristiques qui le précède ne dépend pas de l'extraction de l'arrière plan et que les marqueurs ne sont pas utilisés.
2. La deuxième architecture est basée sur la représentation des actions humaines par l'assemblage des images de l'objet en mouvement et la reconnaissance des actions est réalisée à l'aide les algorithmes de type Deep Learning.

B) Proposition d'une approche de décomposition des activités complexes en séquences d'actions simples préalablement établies. Cela nous permet d'ouvrir de bonnes perspectives de connaître les activités complexes réalisées dans des scènes plus réalistes et complexes.

C) Afin d'analyser la performance des approches proposées, plusieurs tests ont été réalisés dans différents scénarios et divers cas, notamment le saut des trames et l'utilisation d'une résolution vidéo plus faibles.

1.4 Organisation des Chapitres

En outre de ce chapitre, le manuscrit comprend cinq d'autres chapitres organisés comme suit ;

— **Chapitre II Perception visuelle des activités humaines :** Ce chapitre est consacré au système de perception visuelle chez les organismes vivants et les humains en particulier. Nous avons donné un aperçu général sur la perception des mouvements par notre système visuel.

— **Chapitre III Reconnaissance des activités humaines basée vision :** Nous expliquerons dans le troisième chapitre, un état de l'art sur les systèmes automatiques de la reconnaissance des activités humaines basés vision.

— **Chapitre IV Descripteur de mouvement proposé :** Ce chapitre est consacré à la présentation des détails de la construction de notre descripteur proposé à base de flux optique.

— **Chapitre V Processus de classification proposée :** Ce chapitre décrit les différentes méthodes de classification proposées et les résultats expérimentaux obtenus.

— **Chapitre VI Analyse de performance :** Dans ce dernier chapitre nous nous concentrerons sur la clarification de notre méthode d'analyse de performance du descripteur proposé.

— **Conclusions et perspectives :** L'ensemble des éléments développés au cours de cette thèse sont résumés dans cette conclusion. Les perspectives d'applications dans lesquelles s'inscrivent ces travaux sont également présentées parallèlement avec les perspectives d'évolution de cette thématique dans le futur.

1.5 Publications

Sur la base des résultats des recherches durant la préparation de cette thèse, nous avons publié :

1.5.1 Article journal :

— **Ladjailia, A.**, Bouchrika, I., Merouani, H. F., Harrati, N., & Mahfouf, Z.(2019).Human activity recognition via optical flow : decomposing activities into basic actions. Neural Computing and Applications, 1-14. Springer. **Impact Factor : 4.664 (2019)**

— Gnouma, M., **Ladjailia, A.**, Ejbali, R., & Zaied, M. (2019). Stacked sparse autoencoder and history of binary motion image for human activity recognition. Multimedia Tools and Applications, 78(2), 2157-2179.Springer.
Impact Factor : 2.101 (2019)

1.5.2 Communications

— **Ladjailia, A.**, Bouchrika, I., Merouani, H. F., & Harrati, N. (2015). On the use of local motion information for human action recognition via feature selection. In 4th international conference on electrical engineering (ICEE). IEEE.

— **Ladjailia, A.**, Bouchrika, I., Merouani, H. F., & Harrati, N. (2015, December). Automated detection of similar human actions using motion descriptors. In Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2015 16th International Conference on (pp. 398-403). IEEE.

1.5.3 Chapitres

— **Ladjailia, A.**, Bouchrika, I., Harrati, N., & Mahfouf, Z. (2018). Encoding Human Motion for Automated Activity Recognition in Surveillance Applications. In Computer Vision : Concepts, Methodologies, Tools, and Applications (pp. 2042-2064). IGI Global.

2 Perception visuelle des activités humaines

2.1 Introduction

La perception ou l'interprétation des actions produites par les humains est primordiale pour leurs interactions avec l'environnement extérieur et la communication sociale. En effet, interagir de manière appropriée et correcte avec une ou plusieurs personnes demande la perception rapide de la signification des actions effectuées, voire même de la prédiction d'actions susceptibles d'être engendrées dans un proche avenir. Ces compétences en interprétation jouent également un rôle crucial dans les diverses formes de coopération pouvant exister entre plusieurs personnes, notamment en facilitant la coordination d'actions productives, successives ou combinées. Ainsi, la motricité intentionnelle est une source d'information pour améliorer les interactions sociales. Cependant, l'accès à ces informations nécessite un système visuel très puissant qui même lorsque les informations disponibles peuvent être largement épuisées, peut détecter des motrices de petites différences pour comprendre ce qui est observé [17].

L'objectif de ce chapitre est d'étudier et de discuter, sans chercher à être exhaustif, le concept du système de perception visuelle et sa relation avec la perception motrice.

2.2 Perception visuelle

Avant de parler de la perception visuelle, nous pensons d'abord que le concept de perception doit être clarifié au sens général. Le dictionnaire **Larousse** définit le concept "perception" par : «*la perception est un événement cognitif dans lequel un stimulus ou un objet, présent dans l'environnement immédiat d'un individu, lui est représenté dans son activité psychologique interne, en principe de façon consciente ; fonction psychologique qui assure ces perceptions* ». Nous avons trouvé dans la littérature scientifique appartenant aux différents domaines de nombreuses définitions de ce concept. À cet égard, nous nous limitons aux explications fournies par *Schacter et al.* [189], comme suit : la perception est l'identification, l'interprétation et l'organisation des informations sensorielles afin de représenter et de comprendre les informations présentées à l'environnement. Toute perception implique des signaux qui passent par le système nerveux, qui à leur tour résultent de la stimulation physique ou chimique du système sensoriel [84]. Par exemple, la vision implique la lumière frappant la rétine de l'œil, l'odeur est médiée par des molécules d'odeur, et l'ouïe implique des ondes de pression. La perception n'est pas seulement la réception passive de ces signaux, mais elle est aussi modelée par l'apprentissage, la mémoire, l'attente et l'attention du destinataire [88]. La perception peut être divisée en deux processus selon *Bernstein* [16] :

— Le traitement de l'entrée sensorielle, qui transforme ces renseignements de bas niveau en information de haut niveau (p. ex., extrait des formes pour la reconnaissance d'objet) ;

— Le traitement qui est lié aux concepts et aux attentes (ou connaissances) d'une personne, aux mécanismes réparateurs et sélectifs (comme l'attention) qui influencent la perception.

Nous allons nous concentrer uniquement sur certains concepts liés à la perception visuelle. Les paragraphes suivants portent sur ce sujet.

2.2.1 Notion de la perception visuelle

Pour promouvoir ce concept, nous présentons cet exemple afin que le concept de perception visuelle chez l'homme soit clair. Une lecture du texte semble à première vue comme un processus simple : nous dirigeons nos yeux vers les lettres, les voyons et nous savons ce qu'elles veulent dire. Mais en réalité, il s'agit d'un processus extrêmement complexe qui implique une série de structures cérébrales spécialisées dans la perception visuelle et la reconnaissance des différents sous-composants de la vue. Par conséquent, on peut dire que ***la perception est l'interprétation de l'information que nous rapportent nos sens, sur notre environnement.*** Cela montre que l'interprétation que nous faisons dépend de nos processus cognitifs et de nos connaissances.

La perception visuelle peut être définie comme la capacité d'interpréter les informations optiques que nos yeux reçoivent. Le résultat de l'interprétation de cette information par notre cerveau est appelé "***perception visuelle***". En conséquence, la perception visuelle est un processus qui commence par nos yeux et se termine par la construction de la perception visuelle par le cerveau. La figure 2.1 montre les détails de ce processus.

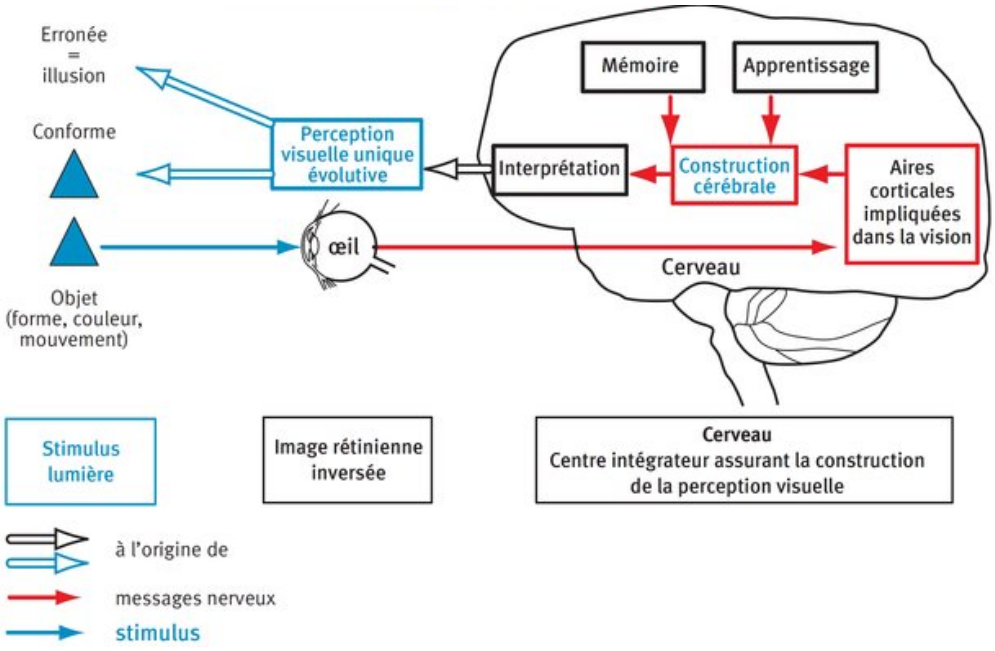


Figure 2.1 – Le schéma de construction de la perception visuelle humaine par le cerveau [58].

Pour illustrer davantage, nous pouvons dire que ce processus se compose de trois étapes suivantes :

— **Photo-réception** : La lumière entre à travers nos pupilles et active des cellules réceptives qui se trouvent dans nos *rétines*.

— **Transmission et traitement basique** : Les signaux produits par ces cellules se transmettent par le nerf optique jusqu'au cerveau. Tout d'abord, il passe par le *chiasma optique* (où les deux nerfs optiques se croisent ; le champ visuel se dirige vers l'*hémisphère gauche*, alors que le champ visuel gauche vers l'*hémisphère droit*) et ensuite l'information prend la relève dans le *noyau géniculé latéral du thalamus* (Figure 2.2).

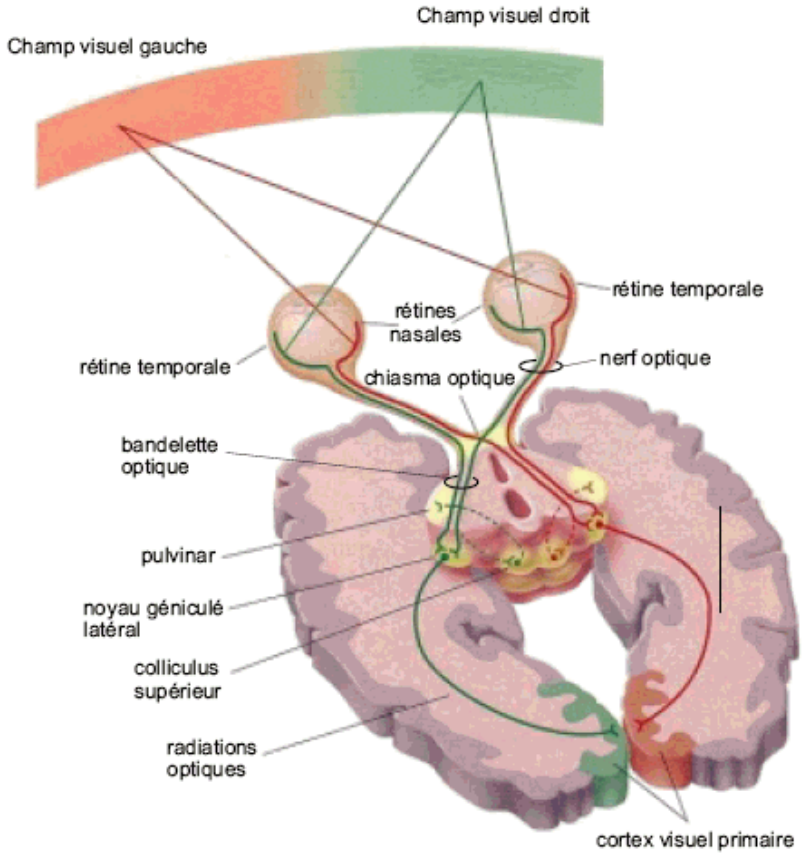


Figure 2.2 – Schéma des voies visuelles. [Source : http://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2002.fort_a&part=57800]

— **Élaboration de l'information et perception :** Finalement, l'information visuelle captée par nos yeux est envoyée au *cortex visuel* du *lobe occipital*.

Au niveau de ces structures cérébrales, l'information est élaborée et envoyée au reste du cerveau pour nous permettre d'interagir avec celle-ci.

2.2.2 Caractéristiques de la perception visuelle

Pour avoir une idée de la complexité du processus de la perception visuelle, nous devons penser ce que fait notre cerveau, quand on voit le ballon de football devant nous (Figure 2.3). La question qui se pose est la suivante : Combien de facteurs devons-nous identifier ?

Figure 2.3 – Complexité de la perception visuelle chez l'homme : la détection en un coup d'oeil du nom de l'objet, la taille, la forme, la position, la couleur, les dimensions, ... [source d'image :Internet]

- **Éclairage et contraste :** Nous voyons qu'il y a des zones noires et blanches et que les lignes ne sont pas très claires à cause de faible éclairage (contraste). Nous pouvons également définir le contour de l'objet et le distinguer de l'arrière plan, malgré leurs couleurs similaires.
- **Taille** : Son diamètre est d'environ 25 cm par rapport aux autres ballons que je connais.
- **Forme** : une forme circulaire.
- **Position** : Le ballon est presque à trois mètre de moi, à ma droite. Facile à atteindre.
- **Couleur** : Il est blanc et noire.
- **Dimensions** : Il est en trois dimensions, c'est une sphère.
- **Mouvement** : Il est immobile, mais susceptible de bouger.
- **Unité** : Il y en a un.

— **Utilité** : Il sert pour jouer au football, on le tire avec le pied.

— **Relation personnelle avec l'objet :** Il est identique à celui que nous utilisons à l'entraînement.

— **Nom** : C'est un ballon de football. Ce dernier processus est appelé *la dénomination.*

De toute évidence, il existe d'autres concepts et des relations entre cet objet et l'observateur qui n'ont pas été mentionnées. Certainement, notre cerveau suit ce processus de manière cohérente et rapide. Par ailleurs, notre cerveau ne perçoit pas l'information de manière passive, en effet, il apporte les connaissances qu'il possède pour compléter ce qu'il perçoit (nous sachions qu'un ballon est sphérique même si nous avons l'impression qu'il est plat sur la photo). Les études de ce contexte montrent qu'il existe trois zones ou lobes du cerveau (occipital, temporel, pariétal) spécialisées pour la perception visuelle. Généralement, une bonne perception va avoir besoin de ces trois zones qui travaillent conjointement.

2.2.3 Perception visuelle en psychologie cognitive

La perception est l'un des grands domaines qui a été étudier par la psychologie. Elle désigne selon la psychologie l'ensemble des mécanismes physiologiques et psychologiques dont la fonction générale est la prise d'information de l'environnement ou de l'organisme lui-même [142]. Ainsi, la perception de point de vue de la psychologie cognitive est une fonction psychique qui permet à l'organisme de capter, d'élaborer et d'interpréter l'information réceptionnée par les organes sensoriels. Il est important de distinguer entre trois éléments dans cet enchaînement : les stimuli qui appartiennent au monde extérieur, les organes sensoriels qui jouent le rôle de récepteurs et la perception qui est un processus psychologique faisant partie du monde intérieur du cerveau. Donc, la perception visuelle est toute sensation résultant d'une impression lumineuse captée(s) par les yeux (la vue). Il est clair que les organes visuels sont presque identiques chez tous les humains à un certain âge, mais la différence réside dans la perception qui dépend clairement d'autres facteurs tels que la mémoire ainsi que les modes de pensée. À titre d'exemple, on identifie une personne en comparant quelques points critiques

de l'impression globale obtenue de cette même personne avec les images internes déjà enregistrées au niveau de la mémoire visuelle. Par conséquent, pour percevoir un objet et surtout l'identifier, vous devez avoir déjà vu des objets similaires. La perception des visages fonctionne depuis la naissance, mais la discrimination entre les visages et leur reconnaissance est une capacité qui est apprise au fil du temps.

2.2.4 Perception visuelle en Neurosciences

Cette section est consacré au système de perception visuelle humaine de point de vue Neurosciences. La figure 2.4 résume le principe de fonctionnement général de ce système. Nous avons expliqué précédemment que ce système repose sur trois piliers : le stimuli, les yeux et la perception au niveau du cerveau. Ce système peut être comparé à une caméra connectée à un ordinateur doté d'un programme d'intelligence artificielle composée d'une base de connaissances et un moteur d'inférence. Les deux figures 2.5 et 2.6 donnent un aperçu sur ce système complexe de la perception visuelle.

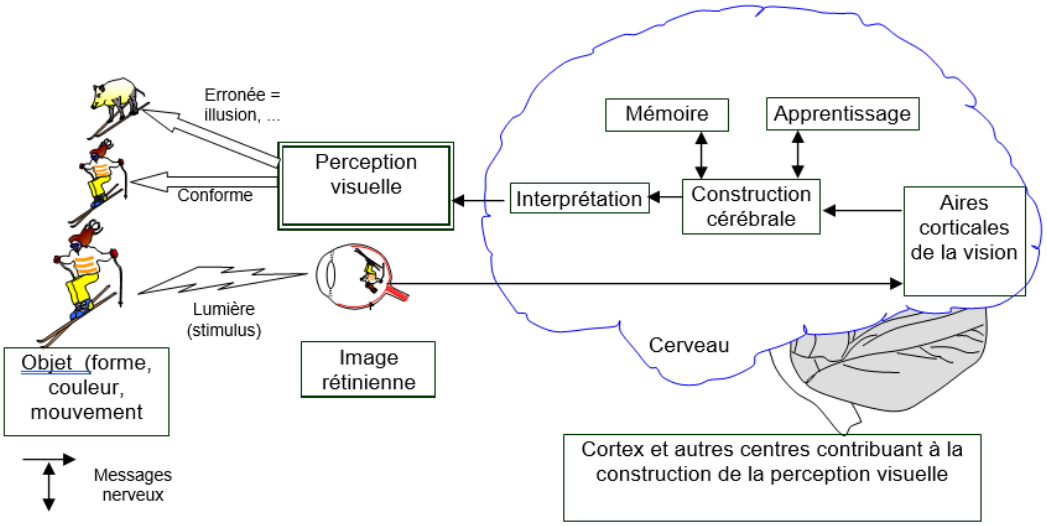


Figure 2.4 – Schéma générale de fonctionnement d'un système visuel chez les hommes [58].

2.2.4.1 Oeil humain

La figure 2.5 montre les composants de base de l'oeil humain :

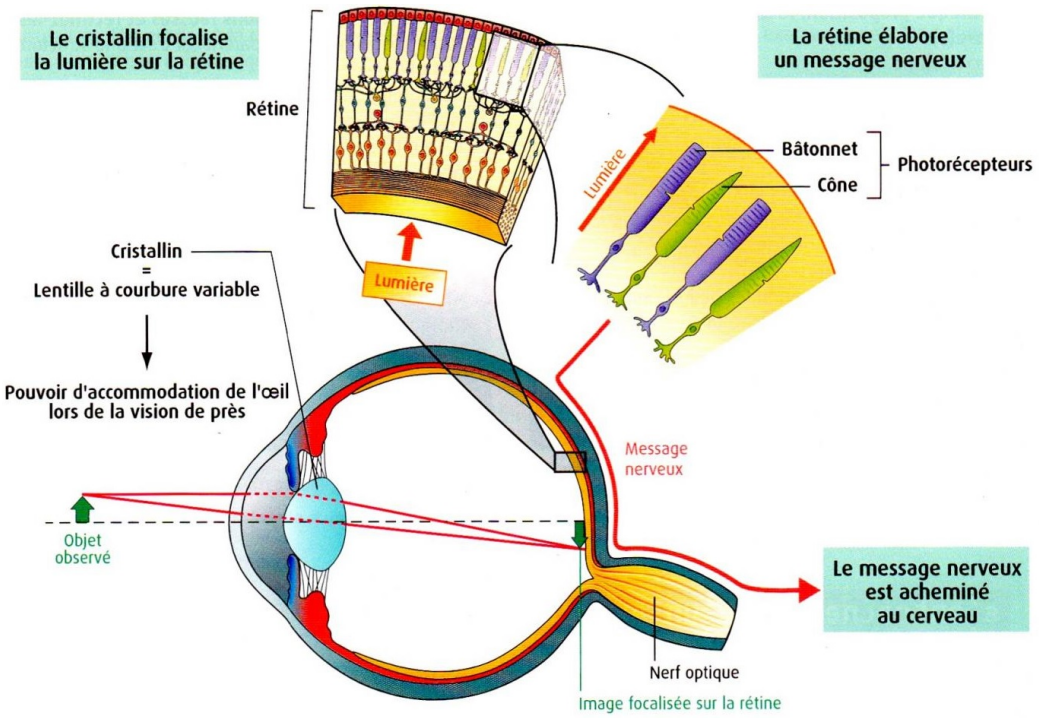


FIGURE 2.5 – Le fonctionnement de l'oeil.[source :http://www.innoverensvt.com/archives/2017/11/12/35859349.html]

— La **cornée** oriente les rayons lumineux vers le centre de l'oeil.

— L'**iris** reçoit la quantité de lumière dont il a besoin en se rétrécissant ou en s'étendant.

— La **rétine** transforme les rayons lumineux en signaux nerveux par excitation physiologique au nerf optique. Elle se constitue de trois couches :

  * La première couche est composée de ***cônes*** et des ***bâtonnets***. Les ***bâtonnets*** permettent de distinguer les lumières de faible intensité (la nuit) et les ***cônes*** permettent la reconnaissance des couleurs (la lumière vive).
  * La seconde couche est le ***fovéa*** ou ***la tâche jaune*** permettant de déterminer le mouvement et le détail des couleurs.
  * La troisième couche est formée des ***cellules ganglionnaires*** terminées par le ***nerf optique***.

2.2.4.2 Aires visuelles dans le cerveau

En neurosciences, il existe de nombreuses manières de connaître des zones du cerveau spécialisées dans le processus de la vision et leur perception. Mais il y a encore d'autres zones inconnues ainsi que leur fonctionnement. Parmi les études faites, on cite *les études cliniques*, par exemple, en cas d'intervention par lésion accidentelle ou pathologique. Dans le même contexte, il est également possible de parler des progrès des *imageries médicales* tel que le TEP (Tomographie par Emission de Positons) et le IRMf (Imagerie par Résonance Magnétique fonctionnelle). Grâce aux méthodes précédentes, on a déduit que le message nerveux visuel émis par la rétine est transporté par les nerfs optiques, puis par d'autres neurones jusqu'à une zone située à l'arrière du cortex occipital de chacun des deux hémisphères cérébraux. Ces deux aires cérébrales forment le cortex visuel primaire (V1). Suite au croisement des nerfs optiques au niveau du chiasma optique, il y a séparation des deux faisceaux de fibres. Chaque aire primaire reçoit des informations provenant des deux yeux. Parallèlement, certaines informations visuelles arrivent directement dans différentes aires visuelles spécialisées dans le traitement de la couleur (V4), des formes (V3) ou encore des mouvements (V5) (Figure 2.6).

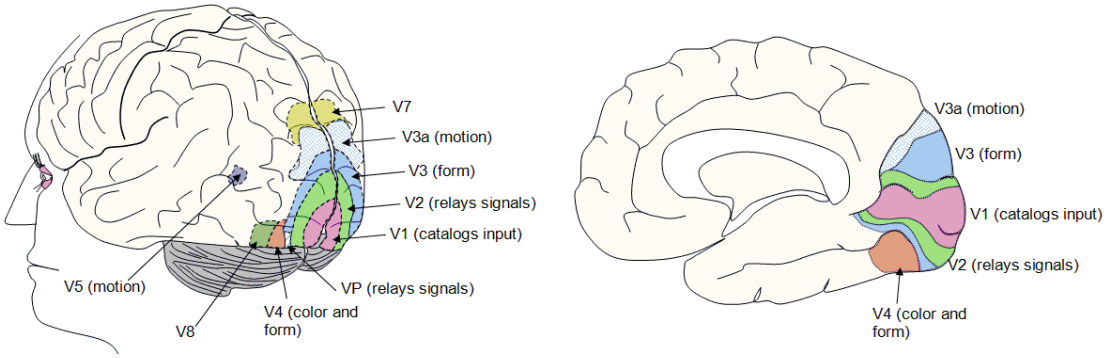


Figure 2.6 – Les détailles d'un système de perception visuelle.[Source :http://lnb.svt.free.fr/doc_1erS/1erS_cours/illu_1S_p3C_chap2.pdf]

2.2.4.3 Perception visuelle : le cerveau intègre les signaux

La vision n'est pas une simple stimulation de la rétine, c'est le cerveau qui traite le message nerveux en provenance de l'œil dans les aires visuelles. Le cerveau fait également interagir ces informations visuelles avec celles des autres aires

spécialisées du cortex : mémoire, langage... etc. Des connexions s'effectuent entre neurones, l'information visuelle est traitée à différents niveaux par plusieurs milliers de neurones ; Ce traitement séquentiel et parallèle de l'information permet aux neurones d'acquérir progressivement depuis les aires visuelles primaires jusqu'aux aires spécialisées, la potentialité de répondre à des stimulations complexes. Le résultat de cette intégration est la perception visuelle de notre environnement avec ses formes, ses couleurs et ses mouvements.

2.3 Perception visuelle et mouvements humains

Les humains peuvent discerner l'état des sujets d'une seule image statique pour en déduire ce qu'ils font, les images cinématographiques fournissent des renseignements encore plus riches et fiables pour la perception des différentes caractéristiques biologiques, sociales et psychologiques de la personne, aussi que les actions et les traits de personnalité du sujet [179]. En revanche, les objets statiques ou immobiles ne sont pas aussi faciles à détecter. Par ailleurs, cette observation a également été illustré par *Darwin* (1872) dans son livre "*The Expression of Emotions in Man and Animals*", où il a écrit : «***les actions parlent plus fort que les images quand il s'agit de comprendre ce que les autres font***». Le système visuel humain est très sensible au mouvement car il tend à concentrer l'attention sur les objets en mouvement. Le mouvement est un événement spatio-temporel défini comme le changement de l'emplacement spatial au fil du temps en fonction de la position de l'observateur. Les auteurs de l'article [57] proposent une définition de ***la perception visuelle du mouvement et la considèrent comme un processus par lequel le système visuel acquiert des connaissances perceptives telles que la vitesse et la direction de l'objet en mouvement.*** Bien que ce processus soit spontané pour le système visuel humain, il est très difficile de reproduire cette capacité sur les systèmes de la vision par ordinateur pour comprendre automatiquement le mouvement humain.

Le domaine de recherche sur la perception visuelle des mouvements humains a été initiées en grande partie par le psychologue suédois *Gunnar Johansson* (1973). Il propose une technique dite ***Moving Lights Display*** ou ***(MLD)*** en bref. Elle

consiste à projeter une séquence de points lumineux en mouvement représentant différents types d'actions humains (marcher, danser, courir). Cette séquence a été obtenue en filmant dans l'obscurité le comportement d'individus ayant des diodes électroluminescences dans les articulations et des parties du corps (Figure 2.7). Divers observateurs sont appelés à voir les acteurs exécutent diverses activités. *Johansson* a tiré de ces travaux une conclusion importante suivante : ***la présentation statique de ces points lumineux ne permet aucune identification de l'action en cours, par ailleurs, il a constaté que la présentation dynamique de ces points même pendant une durée très brève (100 ms) permet une reconnaissance quasi instantanée de l'action en cours et ceci malgré la présence d'un stimulus très appauvri*** [107]. Sur la base de ces expériences, les observateurs peuvent reconnaître différents types de mouvements humains tels que : marcher, sauter, danser et ainsi de suite. De plus, l'observateur peut porter un jugement sur le sexe de l'acteur [125], et même identifier la personne si elle connaît déjà sa démarche [83].

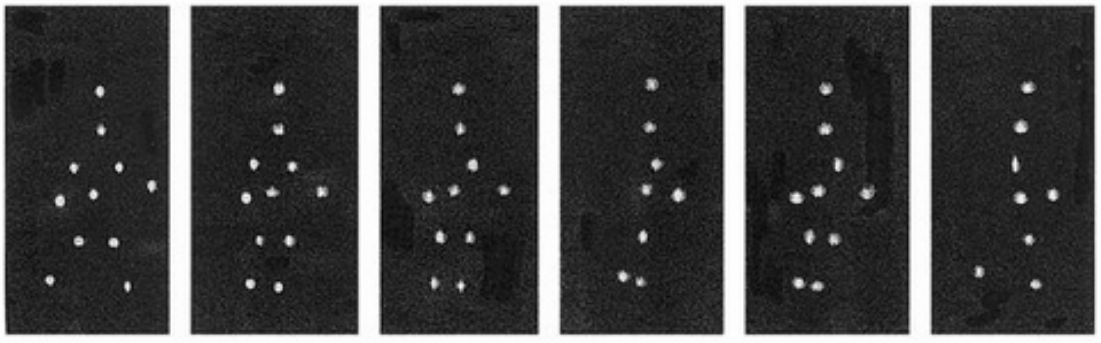


Figure 2.7 – Illustration de la technique ***Moving Lights Display (MLD)*** d'un être humain en marchant. [source de l'image :https://www.biomotionlab.ca/html5-bml-walker]

Cutting et al. [49] ont soutenu que la reconnaissance des mouvements est purement basée sur des caractéristiques dynamiques de la marche par opposition à des études antérieures qui ont été confondus par des indices de familiarité, la taille, la forme ou d'autres sources d'information. Il a atteint un résultat très important et fiable pouvant être utilisé lorsque on a créé un système automatique de la reconnaissance des activités humains, il l'a exprimé comme suit : ***"Bien que les différentes parties du corps humain ne soient pas visibles dans les points et qu'il n'existe aucun lien entre les points lumineux pour montrer la structure squelette du corps humain, l'observateur peut récupérer***

la structure complète de l'objet en mouvement. Par conséquent, le mouvement des articulations contient suffisamment d'informations pour la perception du mouvement humain". Il existe d'autres articles qui supportent cette hypothèse et les plus importants sont cels de [18, 62]. Il existe une mine de recherches qui s'efforce de documenter la capacité du système visuel humain à percevoir le mouvement humain à partir d'un petit nombre de points en mouvement comme le prétendent les premières études médicales de *Johansson, Cutting et Murray*. Néanmoins, le processus de perception sous-jacent est mal compris et il y a toujours un manque de recherche qui explique les principes sous-jacents pour représenter et récupérer le mouvement biologique [205].

Deux théories principales ont été inventées pour la perception du mouvement humain à partir du MLD : ***la théorie structurelle et la théorie basée sur le mouvement*** [34]. La ***théorie structurelle*** est la plus ancienne qui a affirmée qu'à l'étape initiale de la reconnaissance, on récupère la structure 3D de l'information de mouvement observé à partir de MLD, puis la structure récupérée est utilisée pour la reconnaissance des activités humaines. ***La deuxième théorie*** est basée sur le mouvement, cela indique que la reconnaissance est fondée directement sur l'information du mouvement sans récupérer la structure de corps humain à partir de la MLD ; à la place l'information de mouvement est extraite en une séquence des trames. Cette théorie est la base de tous les travaux actuels de la reconnaissance des activités humaines. Nous nous appuyons également sur le même principe dans notre approche proposée.

2.4 Notion de geste, action et activité

En se basant sur les principales études de la littérature du domaine de la reconnaissance des activités humaines, les terminologies "***Geste***", "***Action***" et "***Activité***" sont mentionnées de façon interchangeable et contradictoire, avec un certain chevauchement [175]. La figure 2.8 montre ce chevauchement. Un "***geste***" est défini comme le mouvement élémentaire des parties du corps d'une personne, ce mouvement peut n'avoir aucun sens. L'*Action* peut être défini comme une activité atomique ou un simple mouvement effectué par un sujet dans un court

intervalle de temps de quelques secondes, cela peut inclure, par exemple, plier, s'asseoir et agiter les mains. *Poppe* [175] a expliqué le mot "action primitive" par un mouvement atomique au niveau des membres du corps humain. Une ***activité*** peut être décrite comme une séquence d'actions de base effectuées par un individu ou un groupe de personnes. Les cas d'activités comprennent des actions complexes telles que laisser un sac sans surveillance, agresser un piéton ou serrer la main. Il existe d'autres termes qui sont évidemment très difficiles à énumérer. Beaucoup de concepts indiquent qu'il existe un chevauchent entre elles.

Figure 2.8 – Chevauchement des concepts : Geste, action et activité

Dans cette thèse, on peut donner la définition de l'action comme étant une activité simple réalisée par une personne dans un duré très court, cette action est indivisible, et l'activité comme une séquence d'actions réalisée par une ou plusieurs personnes.

2.5 Conclusion

Après cette étude, il est clair que la perception visuelle chez l'homme est étonnante et inhabituelle : elle fonctionne de manière intuitive, alors que les les systèmes automatique modernes ne peuvent pas la suivre. Nous pouvons confirmer qu'il est difficile de parler des similitudes entre les deux systèmes, surtout si on compare les deux systèmes en termes de qualité et de performance. Dans ce chapitre, nous avons abordé la notion de la perception, la perception visuelle et le mode de fonctionnement du système de perception visuelle humain. Nous avons également étudié la relation entre le système de perception visuelle et le mouvement. Dans le même contexte, nous avons aussi expliqué quelques études récentes qui confirment que pour la reconnaissance de mouvement, il suffit d'extraire certaines propriétés du corps humain en cours de déplacement pour que nous puissions connaître le type de mouvement. Enfin, nous avons expliqué quelques termes du domaine tels que : l'activité, l'action et geste et aussi leurs chevauchements.

Dans les chapitres suivants, la question que nous poserons est la suivante : Pouvons-nous compter sur les sciences modernes et en particulier sur les méthodes de la vision par ordinateur et l'apprentissage automatique pour reproduire ce système sur un ordinateur.

3 Reconnaissance des activités humaines basée vision

3.1 Introduction

Ces dernières années, la reconnaissance automatique de l'activité humaine a attiré beaucoup d'attention dans le domaine de l'analyse et de la reconnaissance de contenu des vidéos en raison des demandes croissantes de nombreuses applications, telles que la surveillance automatique intelligente, les environnements de divertissement, les systèmes de santé, la robotique, les systèmes d'interaction homme-machine, les voitures autonomes, etc. Dans un environnement de surveillance, la détection automatique des activités anormales peut être utilisée pour alerter l'autorité connexe des comportements criminels ou dangereux, comme le signalement automatique d'une personne portant un bagage suspect à l'aéroport ou à la gare. De même, dans un environnement de divertissement, la reconnaissance d'activité peut améliorer l'interaction homme-machine (IHM), telle que la reconnaissance automatique des actions des différents joueurs pendant un match de tennis afin de créer un avatar dans l'ordinateur pour jouer au tennis en faveur du joueur. De plus, dans un système de santé, la reconnaissance automatique des activités des patients facilite les processus de réadaptation. De nombreux efforts de recherche sont rapportés pour différentes applications basées sur la reconnaissance de l'activité humaine, plus spécifiquement, l'activité anormale à la maison [11], le tennis

[114], le football [147], les gestes humaines [180], les activités sportives [148], les interactions humaines [80], la circulation piétonnière [104], la simple action [132], les applications médicales [165]. Dans ce chapitre, la reconnaissance de l'activité humaine à partir des séquences vidéo sera examinée et discutée en détail.

3.2 Architecture générale

En général, un système de reconnaissance des activités humaines est divisé en trois niveaux de représentation, la technique de base de ***bas niveau***, les systèmes de reconnaissance de l'activité humaine de ***niveau intermédiaire*** et le niveau applications de ***haut niveau*** [113]. La figure 3.1 montre l'architecture générale des systèmes automatiques de reconnaissance des activités humaines.

— Le ***premier niveau*** qui s'appelle *le niveau de la technologie de base*. Trois grandes étapes de traitement sont considérées ; la segmentation des objets, l'extraction et la représentation des caractéristiques et enfin les algorithmes de détection et de classification des activités. L'objet humain est d'abord segmenté à partir de la séquence vidéo. Ensuite, les caractéristiques de l'objet humain telles que la forme, la silhouette, les couleurs, les poses et les mouvements du corps sont extraites et représentées par un ensemble de caractéristiques. Par la suite, un algorithme de détection et de classification des activités est appliqué aux entités extraites pour reconnaître les diverses activités humaines.

— Le ***deuxième niveau*** ou *les systèmes de la reconnaissance des activités humaines*. Ce niveau concerne l'étude du comportement des objets en mouvement, y compris la reconnaissance d'une activité individuelle, l'interaction de plusieurs personnes et le comportement de la foule, voire la détection et la reconnaissance d'activités anormales.

— Enfin, le ***troisième niveau*** est le *niveau applicatif*. À ce niveau, on peut mettre beaucoup d'applications réelles telles que les environnements de surveillance, les environnements de divertissement, les systèmes de santé, etc.

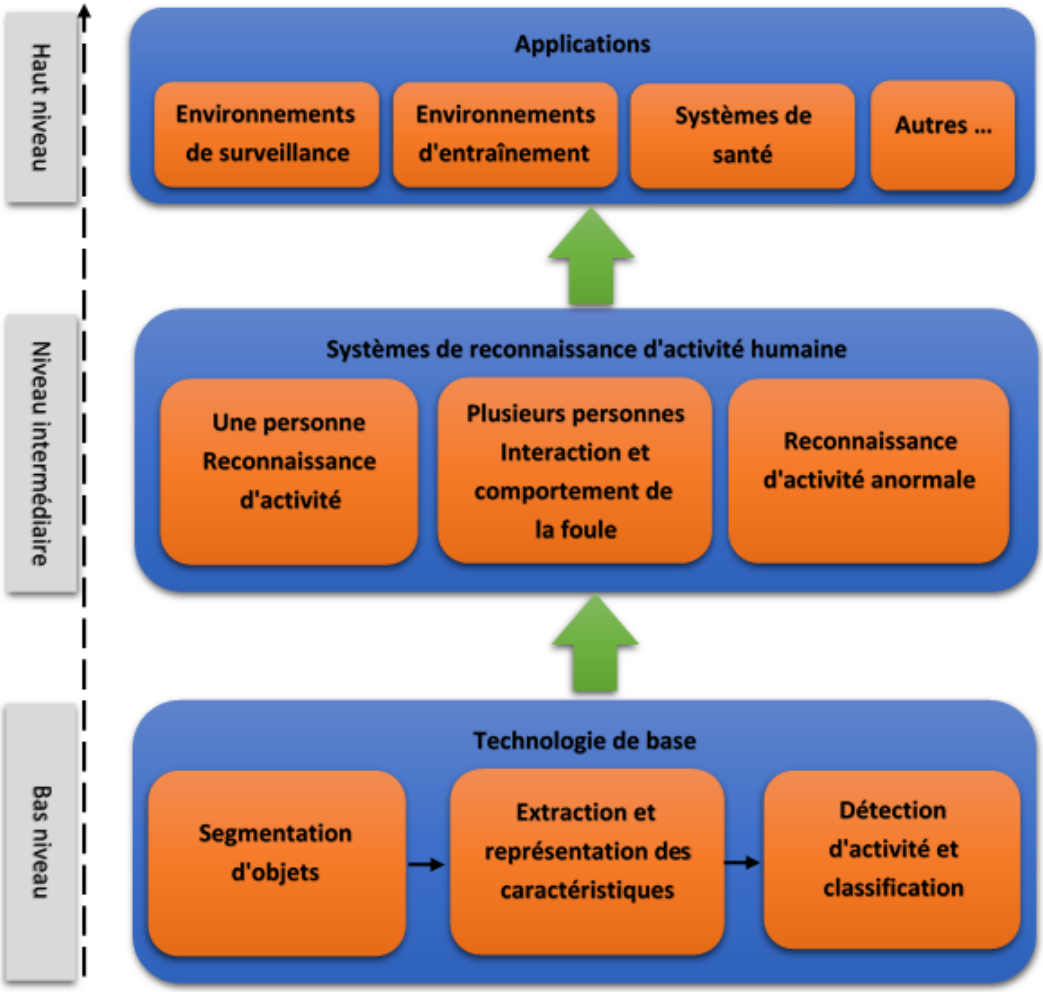


Figure 3.1 – Aperçu général d'un système de la reconnaissance des activités humaines [113].

Nous proposons dans cette thèse une approche générale d'un système de la reconnaissance des actions humaines indépendantes de toutes applications. Le système proposé concerne seulement la reconnaissance et l'analyse des activités d'une seule personne dans une séquence vidéo. Par conséquent, dans cette partie théorique, nous nous concentrerons uniquement sur le premier niveau de l'architecture générale précédente proposé par *Ke et al.* [113].

3.3 Technologie de base

Un système automatique de reconnaissance des activités humaines basé sur la vision par ordinateur est conçu pour extraire seulement les caractéristiques cinématiques du mouvement humain sans avoir besoin d'utiliser des marqueurs ou des capteurs spéciaux. Cela permet de faciliter le processus d'extraction des caractéristiques de mouvement. En fait, nous avons besoin d'une caméra vidéo ordinaire reliée à un logiciel spécial basé sur la vision. Les systèmes de détection de mouvement sans marqueurs sont adaptés aux applications où le montage de capteurs sur le sujet n'est pas une option par exemple dans le cas de la surveillance visuelle. En règle générale, le système se compose de deux éléments principaux :

i) **Plate-forme matérielle** dédiée à l'acquisition de données. Il peut s'agir d'une seule caméra CCTV (Figure 3.2) ou d'un réseau distribué de caméras. Dans notre cas, les vidéos utilisées sont extraites à partir des bases de données standards du domaine afin de vérifier la validité de notre approche proposée.

Figure 3.2 – Caméra de vidéosurveillance (**CCTV camera** : Closed-Circuit TeleVision camera)

ii) **Plate-forme logicielle** dédiée à l'analyse et la reconnaissance des activités humaines. Dans notre cas, cela représente l'implémentions de l'approche proposée.

En ce qui concerne l'architecture de côté logiciel, elle se compose en gros de trois composantes principales, cette division est parfois évidente, mais dans d'autres cas, il y a chevauchement entre les étapes :

i) la détection et le suivi du sujet ou l'étape de segmentation,

ii) l'extraction et la représentation des caractéristiques,

iii) l'étape de classification

La figure 3.3 montre le diagramme de base pour un système de reconnaissance des actions humaines décrivant les différents systèmes sous-jacents. Dans les sections suivantes, nous détaillerons les composantes de base de ce système en se basant sur des études récentes faites dans ce domaine.

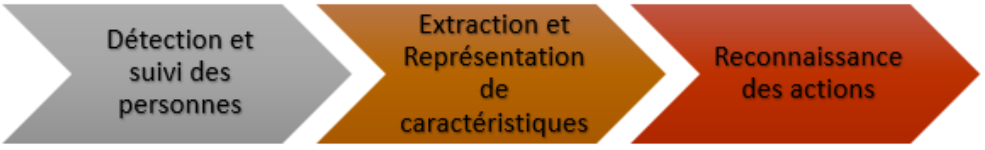


Figure 3.3 – Aperçu générale d'un système de la reconnaissance des actions humaines basé sur la vision par ordinateur.

3.3.1 Détection et suivi de l'objet

La détection et le suivi des personnes sont la première étape importante pour un système automatique de reconnaissance de l'activité humaine.

3.3.1.1 Détection des personnes

Cette étape consiste à segmenter l'image, ce que signifie que les objets humains qui se déplacent doivent être séparés de l'image de l'arrière-plan. En fonction de la mobilité de la caméra, la tâche de segmentation d'objet peut être divisée en deux catégories, la segmentation en mode caméra statique et la segmentation en mode caméra mobile.

— **Caméra statique :** pour ce mode de segmentation, la caméra est fixée dans une position et un angle spécifiques. Puisque le fond ne bouge jamais, il est naturel de construire un modèle de fond à l'avance, de sorte que l'objet de premier plan peut être segmenté à partir de l'image du modèle de fond. On trouvera ci-après certains exemples de ces techniques utilisées pour ce mode.

- **Soustraction d'arrière-plan** : dans ce cas, le modèle de fond ne contient que la scène d'arrière-plan stationnaire sans aucun objet d'avant-plan, et tout changement d'image est supposé être causé uniquement par des objets en mouvement. L'objet d'avant-plan peut donc être obtenu en soustrayant l'image actuelle de l'image d'arrière-plan, suivie d'un seuillage de l'amplitude pour obtenir le masque de segmentation [48, 227, 191].
- **Modèle de Mélange Gaussien ou (GMM)** : chaque pixel est modélisé sous une forme plus compliquée, comme un mélange de Gaussiens, pour s'adapter à différents scénarios de fond. En général, le GMM est appris par l'algorithme de espérance-maximisation. Plus la probabilité d'une valeur de pixel dans le GMM est élevée, plus le pixel appartient probablement à l'arrière-plan [173, 236].
- **Modèle statistique** : en plus de GMM, il existe d'autres méthodes statistique plus compliquées pour modéliser le fond. Par exemple, *Horprasert et al.* [96] ont proposé que chaque pixel ait modélisé par quatre paramètres : la distorsion de luminosité individuelle, la distorsion de chromaticité, la variation de la distorsion de luminosité et la variation de la distorsion de chromaticité.
- **Segmentation par suivi** : *Brendel et al.* [28] ont proposé une méthode de segmentation par le suivi des régions à travers les trames avec un nouvel algorithme de distorsion dynamique circulaire dans le temps (CDTW :circular dynamic-time warping). De plus, *Yu et al.* [237] ont proposé une autre méthode de segmentation en suivant les modèles de mélange Gaussiens de couleur spatiale (SCGMM).

— **Caméra mobile :** Contrairement à la caméra statique à un emplacement fixe et à un angle fixe, la segmentation dans le cas de la caméra mobile (par exemple, la caméra installée sur les voitures, les robots en mouvement, les drones, etc.) est beaucoup plus difficile, parce que deux questions doivent être considérées simultanément, c'est-à-dire le mouvement de l'arrière-plan et le mouvement de chaque objet en mouvement au premier plan. Généralement, la décomposition du mouvement de la caméra est nécessaire pour séparer le mouvement de la caméra du mouvement des objets. Dans les littératures

scientifiques, il existe plusieurs techniques pour résoudre ce problème, nous cite ci-dessous certains exemples de ces techniques.

* **Différence temporelle :** *Murray et al.*[158] ont proposé en premier temps l'estimation de mouvement de la caméra par la méthode de compensation de l'arrière-plan. Par la suite, l'objet en mouvement est détecté en prenant la différence de $t-1$ et t des trames consécutives. Dans le même contexte, *Kim et al.* [116] ont proposé une méthode d'analyse et la compensation du mouvement de la caméra en comparant les caractéristiques des contours entre les images consécutives. Les régions candidates de l'objet en mouvement sont trouvées en différenciant entre l'image transformée t^{me} et l'image $t-1^{me}$. L'objet en mouvement est finalement choisi en combinant l'ensemble des caractéristiques et l'analyse du mouvement.
* **Flux optique :** *Huang et al.* [98] ont proposé un algorithme de détection hybride qui combine les informations de couleur et de flux optique estimer par la méthode Kanade-Lucas-Tomasi (KLT) de la cible. Une architecture de suivi à deux couches est ensuite utilisée pour suivre la cible détectée. Le niveau inférieur utilise la fonction Kanade-Lucas-Tomasi (KLT) qui identifie la correspondance des points locaux entre les images et un filtre au niveau supérieur, qui maintient la relation entre la cible et les points caractéristiques, estime l'état cible suivie.

3.3.1.2 Suivi des personnes

Le suivi du sujet ou tracking en Anglais est un processus de localisation d'un (ou plusieurs) objet en mouvement en temps réel. Il est effectué pour établir la correspondance de la même personne à travers des images consécutives. Les méthodes de suivi sont basées sur les caractéristiques simples de bas niveau tel que la taille de *blob*, l' *aspect-ratio*, la vitesse, la couleur, en plus sur des algorithmes de prédiction pour estimer les paramètres des objets mobiles dans les images suivantes, ceci est basé sur des modèles de mouvement qui décrivent l'évolution des paramètres dans le temps. Les méthodes prédicatives les plus populaires utilisées pour le suivi sont : les méthodes Kalman [19], l'algorithme de condensation [101] et le décalage moyen [45].

3.3.2 Extraction et représentation des caractéristiques

C'est l'étape la plus importante pour les systèmes automatiques d'extraction des caractéristiques sans marqueurs, que ce soit pour l'identification humaine, la classification des activités ou d'autres applications de la vision. En effet, les données cruciales requises pour la phase de classification sont obtenues à ce stade. L'extraction des caractéristiques est le processus d'estimation d'un ensemble de mesures liées soit à la configuration du corps entier, soit à la configuration des différentes parties du corps dans une scène donnée et leur suivi sur une séquence d'images. Les caractéristiques doivent présenter un certain degré de discriminabilité entre les différents groupes d'activités humaines. Différents types de caractéristiques sont utilisés telles que les *trajectoires* des positions articulaires estimées par la récupération de la pose des différentes parties du corps humain [220, 212]. Les caractéristiques basées sur les *contours* sont aussi utilisées dans un certain nombre d'études récentes grâce à l'analyse des données sur les silhouettes [35]. Les caractéristiques *texturales* offrent des résultats prometteurs pour la détection de similarités des actions humaines [239, 154]. *Blank et al.* [21] ont proposé un descripteur basé sur l'analyse des *patches* adjacents en fonction de sa corrélation interne pour comparer les images où il a montré sa puissance pour la détection des actions. Dans le même contexte, *Shechtman et al.* [193] ont utilisé ce descripteur pour étudier la similarité entre les actions. Cependant, la majorité des études considèrent l'utilisation des caractéristiques basées sur le *mouvement* comme une meilleure solution pour comprendre les activités humaines [192]. Selon la façon dont les caractéristiques cinématiques sont représentées en fonction des propriétés spatiales, les caractéristiques estimées à ce niveau peuvent être classées en deux grands types :

— **Caractéristiques globales** : où toute l'image ou la région du corps est considérée en même temps.

— **Caractéristiques locales** : se référer aux caractéristiques qui sont extraites de parties plus petites de l'image.

3.3.3 Classification

La classification est un processus permettant d'attribuer une classe ou une catégorie à chaque objet à classer (processus de l'étiquetage), en se basant sur des données de l'étape de l'extraction et représentation des caractéristiques. Après avoir sélectionné les caractéristiques appropriées à partir de l'image ou de la vidéo, les algorithmes de détection et de classification de l'activité constituent la prochaine étape à envisager pour la reconnaissance de l'activité humaine. À ce stade, une description de haut niveau est produite pour déduire ou confirmer l'identité du sujet. Le processus de classification est normalement précédé par des étapes de pré-traitement telles que la normalisation des données, la sélection des caractéristiques et la réduction de la dimensionnalité de l'espace des caractéristiques au moyen des méthodes statistique. Pour obtenir de bonnes performances de reconnaissance, il est essentiel de choisir un algorithme de classification approprié en utilisant la représentation des caractéristiques sélectionnée. Diverses méthodes de reconnaissance de formes sont employées dans les systèmes de reconnaissance de l'activité humaine basés sur la vision, y compris les*réseaux neurones*, *Support Vecteur Machines (SVM)* et le classifieur *K-Nearest Neighbor (K-NN)*. Récemment, le Deep Learning ou l'apprentissage profond est apparu de manière étonnante dans le domaine de classification et cela pour les bons résultats qu'il a fournis [219].

3.4 Reconnaissance des actions : état de l'art

La reconnaissance d'action, ou reconnaissance d'activité, est le nom de la tâche dont le but est de déterminer l'action ou l'activité d'un individu à partir d'une observation. La reconnaissance des activités humaines est une étape importante dans de nombreuses applications, tels que les interfaces hommes-machines (IHM), les services de santé, les conférences intelligentes, la robotique, la surveillance visuelle automatique et bien d'autres. Il s'agit généralement d'une tâche de classification, où les observations sont testées par rapport à un modèle ou une base de données de catégories d'activités connues pour déterminer quelle activité est effectuée pour cette observation. Il y a plusieurs façons de réaliser une de ces

tâches, avec différents modes d'observation de la scène. Dans le cas de la reconnaissance d'activité basée sur la vision, la tâche consiste à étiqueter les images, qui sont généralement une séquence d'images ou vidéo, avec l'une des étiquettes d'activité qui sont ciblées au préalable.

Généralement, le processus de reconnaissance peut être considéré comme une combinaison de deux problèmes successifs : il faut d'abord trouver une représentation appropriée pour les observations, puis utiliser ces représentations pour la classification afin de déterminer les actions qui ont lieu sur la scène observée. La représentation de l'observation devrait idéalement être invariante par rapport aux propriétés humaines telles que les vêtements et les caractéristiques du corps, le point de vue, les occlusions et les auto-occlusions et autres caractéristiques environnementales telles que l'éclairage et le fond.

Les études (Survey) sur la reconnaissance des activités et les actions humaines basées sur la vision par ordinateur proposent différentes taxonomies et approches [1, 175, 206, 210, 226, 60, 41]. En outre, les méthodes peuvent d'abord être classées en fonction de leurs propriétés spatiales, puis en fonction de leurs propriétés temporelles [226]. En général, ces taxonomies sont basées sur la façon dont les caractéristiques sont extraites et représentées. Les détails de la représentation globale sont illustrés dans la section 3.4.1 et de la représentation locale dans la section 3.4.2 . En raison de la popularité récente des réseaux de neurones de type Deep Learning dans le domaine de la vision, la section 3.4.3 est consacrée aux méthodes de cette approche appliquées aux tâches de reconnaissance des activités. Contrairement aux approches qui sont mentionnées précédemment, les réseaux de neurones profonds ou Deep Learning sont mieux connus pour leur capacité à apprendre automatiquement les caractéristiques distinctives et donc à apprendre à représenter les images elles-mêmes. La section 3.4.4 concerne les approches basées sur les poses pour déterminer l'action humaine.

3.4.1 Approche globale

Pour les représentations globales que l'on appelle parfois *les méthodes holistiques, la région d'intérêt (ROI)* d'une personne est codée dans son intégralité.

Dans la plupart des cas, l'étiquetage ou la détection des parties du corps n'est pas nécessaire, au lieu de cela, les caractéristiques sont calculées de façon dense sur une grille délimitée par région d'intérêt ; le sujet est habituellement dérivé d'une image par application d'une soustraction d'arrière-plan. Le traitement des représentations globales est basé sur des informations de bas niveau provenant de silhouettes, de contours ou de flux optiques [175]. Cependant, ces méthodes sont sensibles aux bruits, aux occlusions et aux variations du point de vue de la caméra. De nombreuses études de recherche ont soutenu que les données de silhouette fournissent de bons indices pour la reconnaissance de l'activité avec l'avantage d'être insensible à la texture, au contraste et aux changements de couleur [226]. Cependant, les méthodes basées sur la silhouette dépendent de la précision de la segmentation de l'arrière-plan qui ne peut être garantie dans les scènes extérieures. Des études récentes ont montré que les silhouettes bruitées peuvent être utilisées pour la reconnaissance de l'activité grâce à l'utilisation de meilleures techniques de concordance (Matching), notamment la distance de Chamfer, la corrélation de phase ou le descripteur contextuel de forme dérivé à partir des données de silhouette [164, 166]. Un autre type important de caractéristiques utilisé pour la représentation globale est *le flux optique* qui est extrait depuis des trames consécutives pour représenter le mouvement pendant que le sujet effectue une activité [2].

Wang et al. [222] ont appliqué la transformée R sur les silhouettes extraites en affirmant que la représentation obtenue est invariable en translation et en échelle. Le principal avantage de la transformation R est son faible coût de calcul ainsi que son invariance géométrique, un ensemble de (HMM : Hidden Markov Model) sont utilisés pour l'entraînement des caractéristiques extraites afin de détecter les activités. *Yamato et al.* [231] ont quantifié des images de silhouette en *super pixels* de telle sorte que chaque pixel indique le rapport entre les pixels noirs et les pixels blancs dans la petite région considérée. *Weinland et al.* [224] ont décrit une représentation compacte et efficace qui repose sur la correspondance d'un ensemble de modèles de pose de repère statiques discriminatoires. La méthode ne tient pas compte de l'ordre temporel des séquences. Dans leur travail, les modèles de silhouette sont comparés aux données de contours à l'aide de la distance de Chamfer et éliminent ainsi le besoin de segmentation de l'arrière-plan. Concernant le flux optique, *Polana et Nelson* [159, 174] ont calculé la texture temporelle

pour reconnaître les événements en fonction de leur mouvement. Pour la reconnaissance de l'activité humaine, les caractéristiques sont basées sur l'amplitude du flux optique contenu dans des cellules non-chevauchantes d'une grille régulière (non-overlapping cells of a regular grid). Dans une autre étude, *Ali et al.* [6] ont dérivé un ensemble de caractéristiques cinématiques du flux optique telles que la divergence, la vitesse, les champs de flux symétriques et anti-symétriques. Une méthode d'apprentissage à instances multiples est utilisée conjointement avec l'analyse en composantes principales pour déterminer les modes cinématiques.

En général, Les approches globales se divisent en deux grands groupes de bases. Les approches qui sont basées sur la construction des volumes spatio-temporels des caractéristiques en espace et en temps à partir de séquences vidéo et les approches séquentielles où la structure et l'analyse d'exécution de mouvements sont considérées.

3.4.1.1 Méthodes volume spatio-temporel

Les approches globales de type volumes spatio-temporels considèrent qu'une vidéo est un empilement 3D des images qui la composent. Cet empilement de silhouettes d'un sujet représente la forme de l'évolution en temps de son mouvement. L'exemple montré dans la figure 3.4 illustre clairement le modèle de mouvement résultant de l'empilement de silhouettes de trois actions simples : courir, marcher et lever les mains respectivement de la base de données Waizmann [86, 20].

Figure 3.4 – Approche globale : volumes spatio-temporels [175]

La reconnaissance de l'action humaine se fait par l'introduction de mesure de similarité entre les volumes obtenus. Dans le même contexte, *Bobick et al.* [22]

proposent une méthodologie de construction de deux types de volumes spatio-temporels : ***Motion Energy Image (MEI)*** qui représente la cumule des pixels en mouvement et ***Motion History Image (MHI)*** qui décrit les zones de forts mouvements de pixels. La figure 3.5 montre un exemple de chacun de ces types.

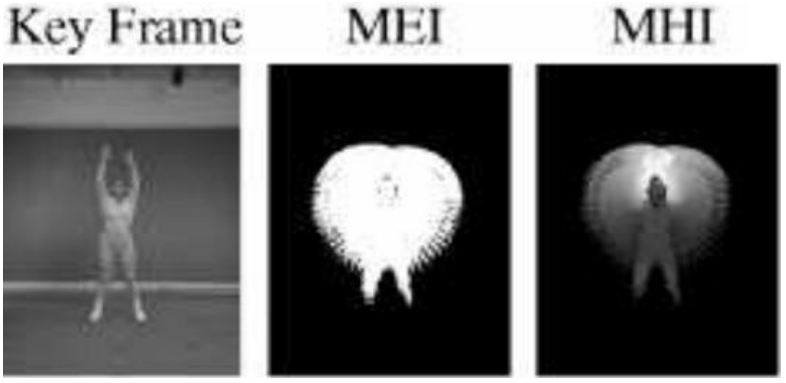


Figure 3.5 – Approche globale séquentielle : à base de MEI (Motion Energy Image) et MHI (Motion History Image) [22]

La combinaison et la concaténation de ces deux composants permettent ensuite la création d'un descripteur de mouvement. L'introduction de moments géométriques [97] augmente la qualité des résultats obtenus.

3.4.1.2 Méthodes séquentielles

La reconnaissance des actions élémentaires par les méthodes séquentielles est basée sur l'analyse d'une séquence d'éléments ordonnés descriptifs, ces derniers sont les parties du corps humain en mouvement dans la séquence vidéo pour l'extraction de silhouettes ou la correspondance (appariements) de squelettes. Ces éléments descriptifs sont extraits au cours du temps afin d'estimer la pose du sujet à chaque image de la séquence. On distingue deux catégories suivantes :

— **Les méthodes basées exemple :** elles reconnaissent les actions élémentaires comme une suite ordonnée et structurée d'exemples représentant celles-ci. On peut citer par exemple l'algorithme "temporel Dynamic Time Warping (DTW)" [15] qui est usuellement employé pour obtenir une invariance en terme de vitesse d'évolution des exemples. La figure 3.6 montre de façon intuitive le recalage entre deux séquences avec une variation de vitesse d'exécution de mouvement.

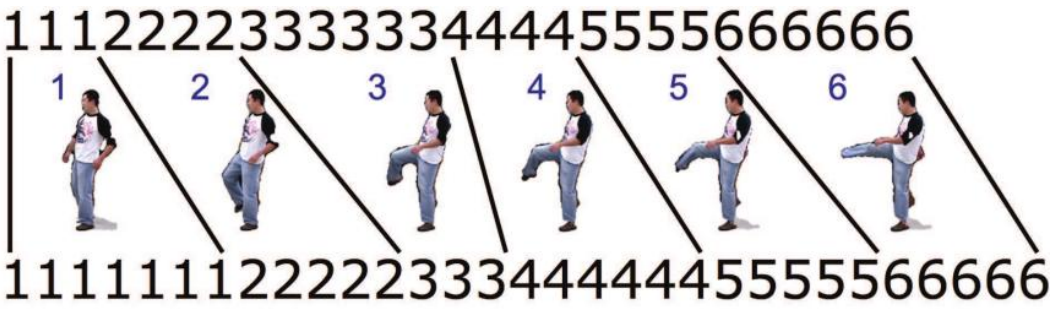


Figure 3.6 – Un algorithme de recalage représentant l'action *Étirer la jambe* entre deux séquences. L'exécution de ces deux séquences est faite avec une variation de vitesse non-linéaire entre elles [1].

— **Les méthodes basées sur les états probabilistes :** ces méthodes ont besoin d'un modèle génératif des probabilités d'observation pour la représentation des actions humaines au cours du temps, la figure 3.7 montre un exemple d'un modèle génératif à base de HMM.

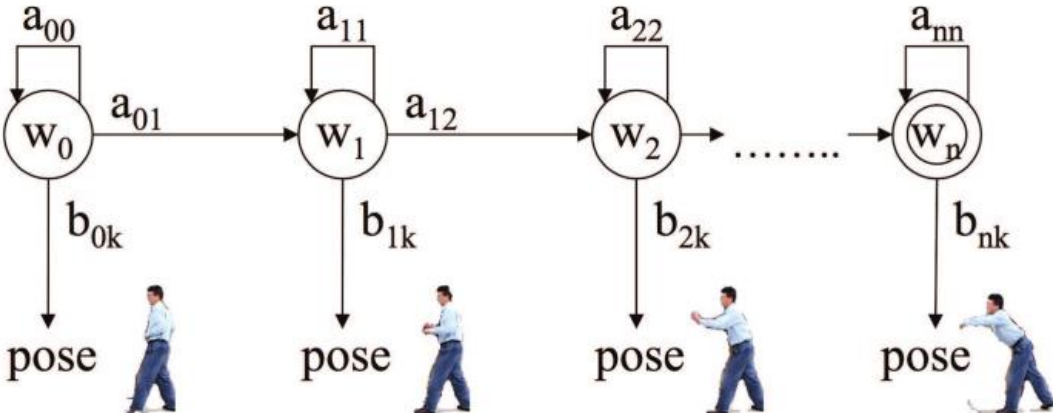


Figure 3.7 – Un modèle génératif à base de HMM pour l'action étiré le bras. Chaque image correspond à une pose k dont la probabilité d'apparition b_{ik} est la plus forte suivant l'état w_i considéré [1].

3.4.2 Approche locale

Pour la reconnaissance d'activité à l'aide de représentations locales, une collection de patchs indépendants à l'intérieur d'une image est analysée pour générer un vecteur de caractéristiques discriminantes pour l'activité observée. Les représentations locales n'ont pas besoin d'une localisation précise ou d'une soustraction

de l'arrière-plan et elles ont l'avantage d'être dans une certaine mesure invariable à la transformation de l'apparence, au bruit de fond et aux occlusion partielle [175]. Les patchs locaux sont décrits par des descripteurs locaux basés sur la grille qui résument localement l'observation dans les cellules de la grille dans le cas de trames fixes. Contrairement à la représentation globale, les caractéristiques locales ne sont pas en relation avec des parties du corps ou des positions spatiales spécifiques d'une image. Les actions ou les activités sont codées en fonction des statistiques des caractéristiques éparses (the sparse features). Le principal avantage de l'utilisation de caractéristiques locales est l'absence de nécessité de détection des personnes ou de localisation des différentes parties du corps [226]. Les descripteurs de points d'intérêt spatio-temporels, qui sont similaires aux points d'intérêt 2D classiques comme SURF et SIFT, sont devenus le type le plus populaire de caractéristiques locales utilisé pour la reconnaissance des actions [134]. La figure 3.8 montre un exemple de la détection de point d'intérêt spatio-temporel lorsque le sujet marche.

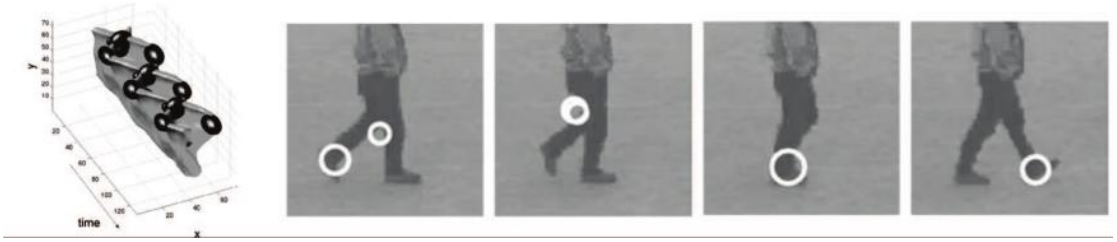


Figure 3.8 – Exemple de la détection d'un point d'intérêt spatio-temporel lors du mouvement de marche [134].

Pour l'utilisation des caractéristiques orientées mouvement pour la reconnaissance d'activité humaine, *Yeffet et al.* [234] ont proposé un descripteur de modèle trinaire local pour coder le mouvement humain à partir d'une séquence d'images. Le nombre trinaire est généré à partir d'un processus de correspondance entre les patchs d'une trame donnée et les patchs adjacents résidant à la fois des trames précédente et suivante respectivement. Le processus de correspondance est basé sur le descripteur d'auto-similarité des textures [193]. Le codage de l'action se fait de la même manière que le codage de l'opérateur binaire local pour décrire le déplacement des patchs entre les trames adjacentes. Un vecteur de caractéristiques basé sur un histogramme est construit à partir de la concaténation qui résulte de la division de l'image en une grille. *Kliper-Gross et al.*, [118] ont utilisé la même approche du modèle de mouvement trinaire local rebaptisé Motion Interchange Pattern (MIP) pour la reconnaissance automatique des activités humaines.

Kliper-Gross et al. ont présenté un mécanisme de suppression afin de séparer les bordures statiques des bordures liées au mouvement. De plus, afin de tenir compte du mouvement de la caméra, la procédure de compensation de mouvement est intégrée dans la description du mouvement local réel basée sur la transformation affine. Pour l'étape de classification, des sacs de caractéristiques visuelles (bag of visual features) sont utilisés avec SVM. *Oshin et al.* [169] ont présenté le descripteur du mouvement relatif pour la reconnaissance de l'activité dans des scénarios non contraints utilisant uniquement le mouvement pour sa création. Le descripteur est basé sur la distribution des points d'intérêt spatio-temporels à l'intérieur de régions localisées.

Le flux optique est également utilisé pour la reconnaissance de l'action humaine par représentation locale des caractéristiques. *Chaudhry et al.* [39] ont argumenté que l'utilisation récente de descripteurs basés sur des histogrammes complexes peut échouer à un point, car ils ne sont pas de nature euclidienne. Un histogramme du flux optique orienté (HOOF :Histogram of oriented optical flow) est proposé avec l'avantage de l'invariance de la direction du mouvement. Les caractéristiques de HOOF sont dérivées à chaque image sans qu'il soit nécessaire de procéder à une segmentation préalable ou à une soustraction d'arrière-plan. Les noyaux de Binet-Cauchy ont été étendus pour permettre l'appariement des histogrammes non-linéaires des séries temporelles [39]. La méthode a été testée sur la base de données Weizmann et qui a rapporté un taux de classification élevé de 95,66%. *Ikizler et al.* [100] ont combiné l'utilisation des bordures d'une figure humaine ajustée par de petits segments de ligne avec des informations de mouvement évaluées par le flux optique. La transformation de Hough est appliquée pour détecter les segments de ligne. La représentation compacte présentée dans son travail a été testée dans différentes conditions difficiles avec une grande précision pour la reconnaissance d'action. La sélection des caractéristiques permet de réduire l'espace original de 108 dimensions en un espace plus petit de 30 caractéristiques. *Martinez et al.* [153] ont calculé le flux optique pour approximer la vitesse pour chaque pixel. Les vecteurs de flux obtenus sont accumulés dans un histogramme par image pondéré par la norme tandis que les orientations de mouvement sont quantifiées dans 32 directions principales. Un descripteur basé sur un histogramme de 192 bacs est obtenu pour chaque action. Les résultats obtenus sur l'ensemble de données de Weizmann montrent que la méthode peut atteindre une précision

moyenne de 95 % en utilisant le classifieur de machine vectoriel de support. Dans une autre étude de *Ladjailia et al.* [131], les auteurs proposent une approche pour coder une séquence de trames dans un vecteur de caractéristiques décrivant l'action réalisée par une personne. La méthode ne dépend pas de la soustraction de l'arrière-plan pour la détermination des caractéristiques du mouvement. Cela s'explique par le fait qu'il est coûteux et complexe de déployer la soustraction du bruit de fond pour les applications de surveillance en temps réel en raison du processus de mise à jour du modèle de bruit de fond qui est influencé par plusieurs facteurs tels que la fouille de fond, les conditions météorologiques et autres effets environnementaux extérieurs. Il a été inspiré par les travaux de *Kliper-Gross et al.* [118] afin de proposer un modèle d'échange de mouvement pour la reconnaissance des actions. Puisque les descripteurs locaux sont connus pour leur efficacité et leur robustesse pour le codage des textures à des fins de reconnaissance, il a proposé un descripteur qui est basé sur la construction d'une caractéristique qui reflète le déplacement du patch d'une image à l'autre.

3.4.3 Approche à base des poses

Dans le cas de la reconnaissance de l'action à l'aide d'une représentation basée sur la pose, les parties du corps humain sont d'abord récupérées ou reconstruites à l'aide de modèles spécifiques. Bien que les approches basées sur des modèles soient généralement complexes et exigent des coûts de calcul élevés, ces approches sont les plus populaires pour l'analyse du mouvement humain en raison de leurs avantages [230]. Le modèle peut être un modèle structurel en 2 ou 3 dimensions, un modèle de mouvement ou un modèle hybride. Le modèle structurel décrit la topologie des parties du corps humain comme la tête, le torse, la hanche, le genou et la cheville par des mesures telles que la longueur, la largeur et la position. Ce modèle peut être constitué de formes primitives basées sur la mise en correspondance entre des caractéristiques de bas niveau comme les bords. Les modèles en stick et volumétrique sont les méthodes structurelles les plus couramment utilisées. *Akita* [4] a proposé un modèle composé de six segments comprenant deux bras, deux jambes, le torse et la tête. *Guo et al.*[89] ont représenté la structure du corps humain par un modèle en forme de stick qui avait dix bras articulés reliés par six joints. *Rohr*

[185] a proposé un modèle volumétrique pour l'analyse du mouvement humain en utilisant 14 cylindres elliptiques pour modéliser le corps humain. *Karaulova et al.* [110] ont utilisé la figure du stick pour construire un modèle hiérarchique de la dynamique humaine en utilisant des modèles de Markov cachés (HMMs). *Gavrila et al.* [76] ont décrit un modèle 3D de récupération de pose basé sur une recherche d'images de synthèse contre des images réelles en utilisant la distance de Chamfrein pour différentes vues. Le principal mérite de l'utilisation de modèles 3D est l'invariance du point de vue à condition que l'estimation de la pose soit faite avec précision [225].

Wang et al. [216] ont montré que la reconnaissance des actions commence par une étape de détection de sujet par trame pour initialiser l'espace de recherche des parties locales du corps humain, puis l'intégration des parties détectées dans la structure cinématique humaine par un modèle graphique structurel en arbre. La configuration finale de l'articulation humaine est finalement utilisée pour déduire la classe d'action exécutée en fonction du comportement de chaque partie individuelle et de la variation globale de la structure (Figure 3.9).

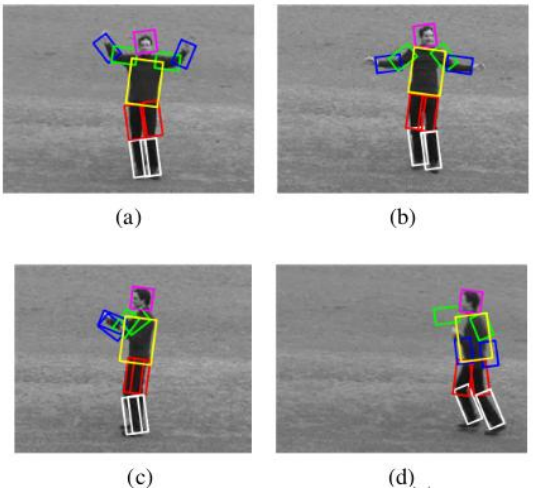


Figure 3.9 – Estimation de pose pour obtenir des configurations pour différentes actions humaines :(a) Hand Waving ; (b) Hand Clapping ; (c) Boxing ; (d) Jogging [216].

Récemment, des chercheurs montrent que les deux domaines ; la reconnaissance des actions humaines et l'estimation des poses sont étroitement liées. Dans ce contextes *Luvizon et al.* [150] ont proposé un framework multitâche à base de Deep Learning pour l'estimation des poses 2D et 3D à partir d'images fixes et pour

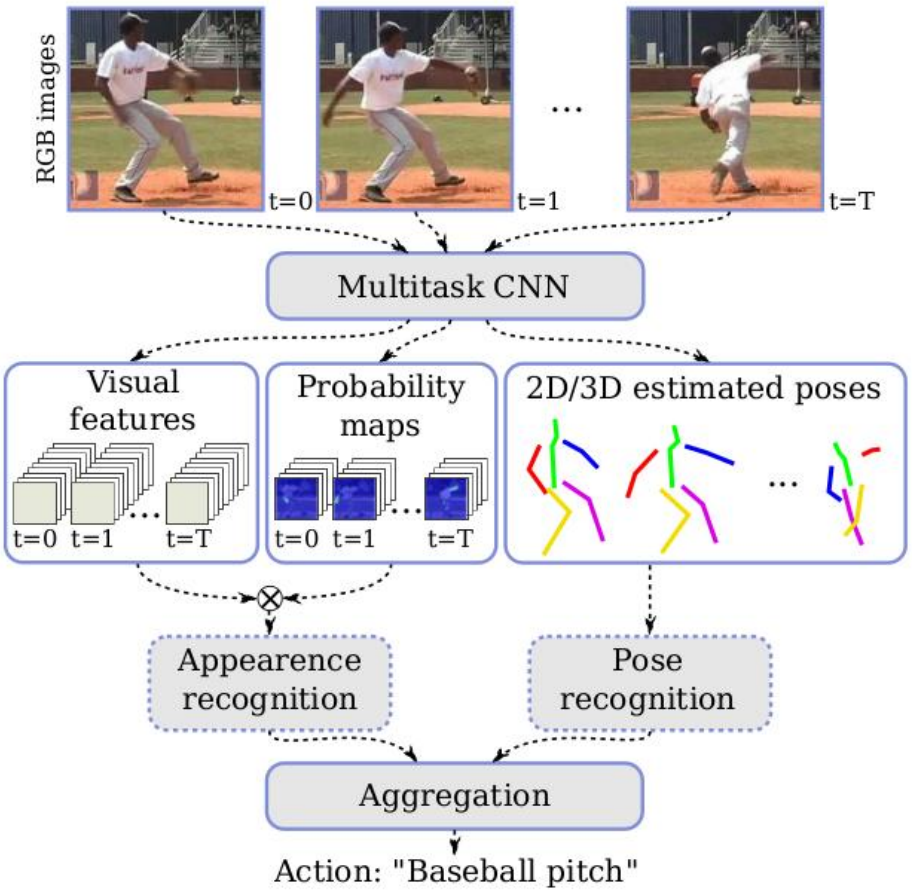


Figure 3.10 – Approche multitâche pour la reconnaissance des actions et l'estimation de la pose [150].

la reconnaissance des actions humaines à partir de séquences vidéo. Le diagramme de la figure 3.10 montre les étapes de la construction de ce modèle.

3.4.4 Approche basée Deep learning

Récemment, le Deep Learning domine plusieurs domaines d'application, en particulier le domaine de la détection des objets et la reconnaissance des actions. Par conséquent, dans le paragraphe suivant, nous illustrons un bref aperçu sur les principes de base de Deep Learning, et nous clarifierons ainsi la relation entre celui-ci et le domaine de la reconnaissance des activités humaines.

3.4.4.1 Deep Learning

Les tâches traditionnelles des modèles de la reconnaissance reposent sur l'extraction manuelle des caractéristiques, que ce soit le domaine de la vision par ordinateur, la reconnaissance de la parole ou le traitement de la langue. Les approches à base des descripteurs HOF dans *Barron et al.* [12] , SIFT dans *Lowe* [146] et HOG dans *Dalal et al.*[50] utilisent des caractéristiques de bas niveau ; ainsi que, de moyenne niveau comme K-means, Sparse coding ou Bag-Of-Words, ils ont besoin d'une étape finale de classification ou de régression. *L'apprentissage en profondeur ou Deep learning* est un terme récemment apparu et il est souvent utilisé de manière interchangeable pour les réseaux de neurones profonds, les réseaux de neurones récurrents, les réseaux de neurones convolutionnelles (CNN), les machines de Zoltzmann profondes, etc. Le terme a attiré une attention remarquable puisque certaines méthodes basées sur l'apprentissage approfondi [54, 126, 136, 64, 111] ont prouvé leur efficacité, surtout lorsqu'on parle de Big data ou une grande quantité de données d'entraînement et de matériel approprié, des unités de traitement graphique appropriées (GPU) sont disponibles. Depuis 2016, l'apprentissage profond domine le domaine de la vision par ordinateur et les conférences de haut niveau tel que CVPR (Computer Vision and Pattern Recognition). La principale innovation derrière l'apprentissage en profondeur est qu'il n'exige pas des définitions explicites des caractéristiques, mais l'extraction de celles-ci se fait de manière automatique à partir des données brutes et de manière hiérarchique de bas niveau jusqu'au haut niveau. Par exemple, la résolution du problème de la reconnaissance d'objets à l'aide de Deep Learning se fait généralement à l'aide de plusieurs couches où les pixels, les bords, les motifs, les parties de corps sont conceptualisés de façon hiérarchique (Figure 3.11).

Deng et al. [55] ont affirmé que l'apprentissage profond ou Deep Learning est une classe d'algorithmes d'apprentissage qui :

— Utilise une cascade de couches multiples d'unités de traitement non-linéaire pour l'extraction et la transformation des caractéristiques. Chaque couche utilise la sortie de la couche précédente comme entrée.

— Apprendre de façon supervisée (P. ex. Classification) et/ou non supervisée (p. ex., analyse de modèles).

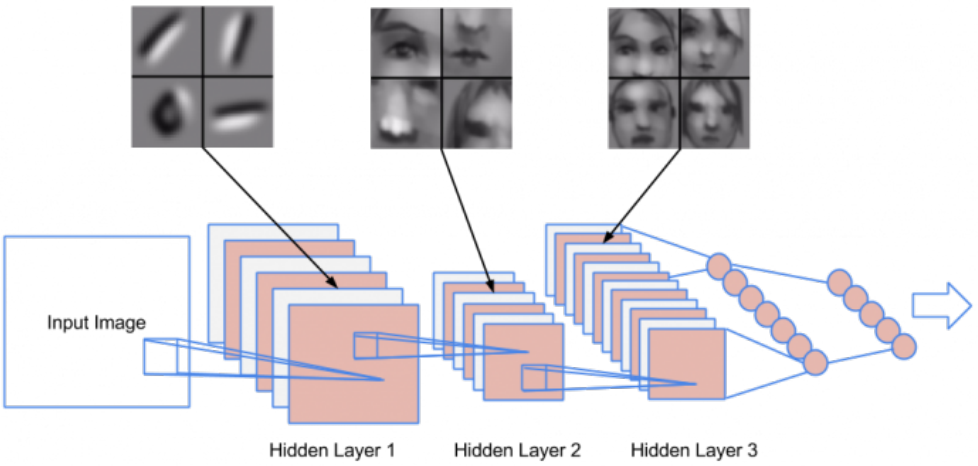


Figure 3.11 – Architecture de Deep learning de type CNN. [source de l'image : https://www.deeplearningitalia.com/analysis-of-deep-learning-models-using-deep-echo-state-networks-deepesns/]

— Apprendre de multiples niveaux de représentations qui correspondent à différents niveaux d'abstraction ; les niveaux forment une hiérarchie de concepts.

3.4.4.2 Deep Learning et la reconnaissance des activités humaines

Le Deep Learning a été appliqué avec succès à de nombreux problèmes au cours de ces dernières années ; et le domaine de la reconnaissance des activités humaines et l'un d'entre eux. La figure 3.12 illustre une taxonomie des principaux travaux de reconnaissance d'actions et de gestes à l'aide des méthodes d'apprentissage en profondeur [9].

Le défi le plus crucial dans le domaine de la reconnaissance des actions humaines par le deep learning est de savoir comment traiter la dimension temporelle. Sur cette base, *Asadi et al.* [9] classent ces approches en trois groupes différents :

— **Modèles 3D :** le premier groupe utilise des filtres 3D dans la couche convolutionnelle. La convolution 3D et le pooling 3D dans les couches CNN permettent de capturer des caractéristiques discriminantes sur les dimensions spatiales et temporelles tout en conservant une certaine structure temporelle.

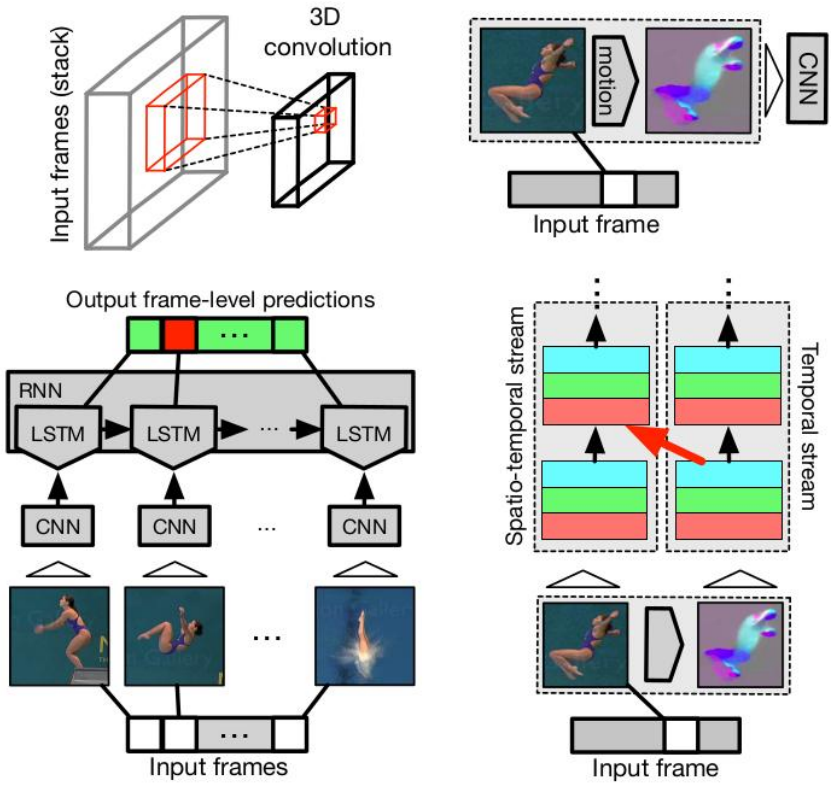


Figure 3.12 – Les différentes architectures et stratégies de fusion. **En haut à gauche** : convolution 3D. En **haut à droite** : pré-calcul du mouvement. En **bas à gauche** : modélisation séquentielle via LSTM. En **bas à droite** : fusion dans un flux spatio-temporel. (figure tiré de [9])

— **Caractéristiques d'entrée basées sur le mouvement :** dans ce cas, les caractéristiques de mouvement sont pré-calculées et entrées sur le réseau sous forme de flux optiques denses 2D.

— **Méthodes temporelles :** le troisième groupe combine le CNN 2D (ou 3D) appliquée à des trames individuelles (ou à des piles de trames) avec une modélisation de séquence temporelle.

3.4.4.3 Typologies

Il existe plusieurs types de Deep learning. Selon le livre "Deep Learning : A Practitioner's Approach" de *J.Patterson et A. Gibson* [170], il existe quatre principales architectures des réseaux profonds suivants :

— **Unsupervised Pretrained Networks (UPNs)** : dans ce groupe, on trouve les architectures suivantes :

  — ***Autoencoders :*** le rôle de ce type est d'apprendre les représentations compressées des ensembles de données. Généralement, ils les utilisent pour réduire la dimensionnalité d'un ensemble de données. La sortie du réseau est une reconstruction des données d'entrée sous la forme la plus efficace.

  — ***Deep Belief Networks (DBNs) :*** Les DBN sont composés de couches de machines de Boltzmann à restriction (RBM) pour la phase de pré-traitement, puis d'un réseau de feed-forward pour la phase de fine-tune.

  — ***Generative Adversarial Networks (GANs) :*** Les GAN sont très habiles à synthétiser de nouvelles images basées sur d'autres images d'entraînement.

— **Convolutional Neural Networks (CNNs) :** l'objectif d'un CNN est d'apprendre des caractéristiques d'ordre supérieur des données par l'intermédiaire de convolutions. Ils sont bien adaptés à la reconnaissance d'objets à partir des images.

— **Recurrent Neural Networks** : les réseaux neuronaux récurrents font partie de la famille des réseaux neuronaux feed-forward. Ils se distinguent des autres réseaux de feed-forward par leur capacité à transmettre des informations au fil du temps. Les réseaux LSTM sont la variante la plus couramment utilisée des réseaux neuronaux récurrents, ils ont été introduits en 1997 par *Hochreiter et Schmidhuber* [93].

— **Recursive Neural Networks** : les réseaux neurones récursifs, comme les réseaux neuronaux récurrents, peuvent traiter des entrées de longueur variable. La principale différence est que les réseaux neurones récurrents ont la capacité de modéliser les structures hiérarchiques dans l'ensemble de données d'apprentissage.

3.4.4.4 Domaines d'applications

l'apprentissage profond s'applique à divers secteurs ou domaines, notamment :

- la reconnaissance visuelle [43, 25]
- la reconnaissance vocale [31]) ;
- la robotique[138] ;
- la bioinformatique [85] ;
- la reconnaissance de formes [92] ;
- la sécurité [56] ;
- la santé [137, 181] ;
- la pédagogie assistée par ordinateur [143] ;
- traduction automatique [10] ;
- ....

Chaque jour, les chercheurs nous proposent un nouveau domaine dans lequel l'apprentissage approfondie peut être appliqué et, chose surprenante que les résultats obtenus sont toujours très satisfaisants.

3.5 Les difficultés et les défis

Un être humain peut facilement et intuitivement connaître le contenu des images (Par exemple la figure 3.13 illustre un chat et un enfant), mais tous les ordinateurs considèrent ces images comme des grandes matrices des pixels (Figure 3.14), on ne peut effectuer que des opérations arithmétiques et logiques sur ces matrices. Ce problème peut être s'exprimer par la notion de l'*écart sémantique* ou (*Semantic Gap*) en anglais [91].

L'écart sémantique est la différence entre la perception de l'image par l'être humain et sa représentation afin que l'ordinateur puisse comprendre son contenu. Le problème de l'écart sémantique n'est pas le seul qui apparaît lors de l'étude du domaine de la reconnaissance des activités humaines, mais il existe plusieurs défis, nous pouvons les classer en catégories suivantes :

(a) Chat (b) Bébé

Figure 3.13 – Ce que l'homme voit.

151	121	1	93	165	204	14	214	28	235
62	67	17	234	27	1	221	37	189	141
20	168	155	113	178	228	25	130	139	221
236	136	158	230	10	5	165	17	30	155
174	148	93	70	95	106	151	10	160	214
103	126	58	16	138	136	98	202	42	233
235	103	52	37	94	104	173	86	223	113
212	15	179	139	48	232	194	46	174	37
119	81	241	172	95	170	29	210	22	194
129	19	33	253	229	5	152	233	52	44
88	200	194	185	140	200	223	190	164	102
113	16	220	215	143	104	247	29	97	203
9	210	102	246	75	9	158	104	184	129
124	52	76	148	249	107	65	216	187	181
6	251	52	208	46	65	185	38	77	240
150	194	28	206	148	197	208	28	74	93
33	183	248	153	168	205	146	100	254	218
130	53	128	212	61	226	201	110	140	183
165	246	22	102	151	213	40	138	8	93
152	251	101	230	23	162	70	238	75	24
187	105	152	83	167	98	125	180	136	121
139	197	55	209	28	124	208	208	104	40
123	19	144	223	62	253	202	108	47	242
220	144	31	16	136	123	227	62	183	163

(a) Matrice 1

29	142	142	75	22	109	111	28	6	5
137	168	41	206	100	70	219	127	114	191
205	154	226	14	89	86	242	67	203	15
247	47	128	123	253	229	181	251	232	28
68	75	24	99	93	63	215	222	102	180
206	246	85	103	215	3	62	64	77	216
126	80	165	149	196	75	186	60	179	193
44	253	164	253	14	216	175	30	46	254
137	23	33	203	241	21	144	63	244	188
32	214	142	121	249	109	99	232	183	71
45	36	152	27	190	137	61	1	237	247
1	14	241	70	2	30	151	67	169	205
32	80	102	32	99	169	91	166	73	214
186	219	9	203	209	240	40	249	119	122
177	252	38	203	119	0	217	139	139	157
154	145	49	251	150	185	235	23	230	156
157	168	223	60	247	118	5	180	16	206
102	208	195	246	140	138	54	191	139	79
17	233	85	169	166	24	49	40	160	97
84	242	247	144	203	3	19	24	198	88
67	67	185	98	123	106	168	105	127	153
37	113	214	252	203	80	146	211	7	16
142	241	66	86	214	133	146	253	189	200
67	215	174	111	189	54	144	56	59	163

(b) Matrice 2

Figure 3.14 – L'image entre le point de vue de l'homme et de l'ordinateur

— **Variation d'échelle :** c'est la capacité que le système connaîtra le même objet de loin ou de proche.

— **Variation de point de vue :** dans le cas de la variation de point de vue, un objet peut être orienté/rotatif dans de multiples dimensions par rapport à la façon dont l'objet est photographié et capturé.

— **Déformation :** l'une des variations les plus difficiles à prendre en compte est la déformation des objets cibles et surtout le suivi de ces objets dans les différentes trames.

— **Occlusions :** la classification de l'image devrait également être capable de gérer les occlusions, où de grandes parties de l'objet que nous voulons classer sont cachées de la vue dans l'image.

— **Éclairage :** le changement d'éclairage est aussi difficile à gérer que les déformations et les occlusions mentionnées ci-dessus.

— **Bruit de fond :** il faut tenir compte du bruit engendré par la font de l'image.

— **Variation intra-classe :** Enfin, les algorithmes de classification des images doivent pouvoir classer toutes ces variations intra-classes correctement.

3.6 Applications

La recherche sur la reconnaissance automatique des activités humaines est alimentée par une large gamme des applications où l'analyse du mouvement humain est déployée, telles que la surveillance automatique intelligente , la biométrie comportementale, les interactions homme-machine, l'animation et la synthèse d'images, nous trouvons aussi l'arbitrage et l'analyse sportifs, analyse des activités anormales, analyse des flux des piétons et des véhicules,etc. Dans cette section, nous choisirons certains de ces champs d'application pour les expliquer et donner plus de détails

3.6.1 Surveillance automatique intelligente

Traditionnellement, il est impossible pour les opérateurs humains de travailler simultanément sur différents écrans vidéo afin de suivre et d'identifier les personnes d'intérêt et d'analyser leurs comportements à différents endroits. Ainsi, il est devenu indispensable pour les scientifiques de la communauté de la vision par ordinateur d'étudier des alternatives visuelles pour automatiser le processus de reconnaissance de l'activité humaine sur différentes vues. Récemment, diverses approches ont été publiées dans la littérature pour accomplir cette tâche basée sur l'utilisation des caractéristiques de base telles que la forme ou la couleur des informations. Toutefois, leur utilisation pratique dans des applications réelles est très limitée en raison de la nature complexe de ce problème [27, 24, 210, 129]. En fait, en raison de l'incapacité des opérateurs humains à surveiller le grand nombre de CCTV (*Closed-Circuit television*) installés dans des zones très sensibles et peuplées

comme les bâtiments gouvernementaux, les aéroports, les centres commerciaux et aissi les universités, ces systèmes sont devenus inutilisables en raison de leur incapacité à être utilisés. Selon la *British Security Industry Association*, le nombre de caméras de surveillance déployées au Royaume-Uni était estimé à plus de 5 millions en 2015 ; ce chiffre devrait augmenter rapidement, en particulier après les attentats terroristes dont un certain nombre de villes d'Europe ont été témoins. Malgré l'énorme augmentation du nombre de systèmes de surveillance, la question de savoir si les systèmes de surveillance actuels ont un effet dissuasif sur la criminalité est encore discutable [27]. Les systèmes de sécurité devraient non seulement être en mesure de prévoir quand un crime est sur le point de se produire, mais, plus important encore, grâce à la détection précoce des personnes suspectes qui peuvent constituer une menace pour la sécurité, le système serait en mesure de décourager les crimes futurs, car il est essentiel d'identifier l'auteur d'un crime dès que possible afin de prévenir de nouvelles infractions et permettre que justice soit faite. De plus, l'utilisation de la technologie de surveillance visuelle intelligente a un large éventail d'applications potentielles en plus de l'analyse du comportement comme le contrôle d'accès, l'analyse des flux de foule et l'analyse de congestion [119, 210].

3.6.2 Interaction homme-machine

L'interaction gestuelle fait de plus en plus partie intégrante des nouveaux systèmes pour les téléviseurs intelligents et les consoles de jeux. Les repères visuels (The visual cues) sont le mode de communication non-verbale le plus important, leur emploi efficace offre aux utilisateurs des moyens prometteurs et novateurs d'interagir avec les ordinateurs. Cela peut même contribuer à améliorer le niveau d'accessibilité et de convivialité pour les personnes ayant des besoins et des exigences spécifiques. Comme dans de nombreux films de science-fiction où l'acteur peut interagir avec les systèmes informatiques en bougeant les mains et en tapant les doigts dans l'air. C'est maintenant une réalité avec l'introduction de *Microsoft Kinect* et les prix abordables des détecteurs de profondeur qui a provoqué le développement rapide et brusque des interactions gestuelles depuis la création de produits commerciaux à une myriade de projets de recherche [183]. Les joueurs de

jeu au lieu d'utiliser des coussinets ou des joysticks, ils peuvent utiliser tout leur corps, leurs mains et leurs jambes comme méthode d'entrée pour contrôler le jeu sans porter de capteurs ou de marqueurs spéciaux. De plus, de nombreux appareils électroniques grand public tels que les télévisions intelligentes ont été développés avec la capacité de laisser les utilisateurs interagir en utilisant des gestes manuels pour échanger entre différents canaux ou contrôler le niveau de volume. Divers Frameworks de développement et outils de programmation sont proposés pour faciliter le processus d'interaction gestuelle à l'aide de Kinect et d'autres capteurs [59, 199]. De plus, il existe un axe de recherche pour créer des environnements interactifs tels que les salles intelligentes qui peuvent réagir à divers gestes humains [128].

3.6.3 Indexation et recherche des vidéos

Avec la croissance constante des sites de partage de vidéos sur YouTube et le téléchargement de plusieurs gigaoctets de contenu multimédia chaque jour, il devient nécessaire de développer des moyens efficaces pour indexer et récupérer les données vidéos au-delà de l'utilisation de simples informations textuelles et tags. Ceci peut être réalisé grâce à des attributs sémantiques qui peuvent être extraits du contenu réel des données vidéo. La résumée vidéo basée sur le contenu a gagné un grand intérêt avec les progrès de la récupération d'images basée sur le contenu [186]. La plupart des premières méthodes utilisent des traits sémantiques simples comme les couleurs et les formes de base pour la recherche des vidéos. Les efforts de recherche récents sont orientés vers la détection d'objets à l'aide de diverses approches, notamment l'utilisation du sac visuel des mots (visual bag of words). Il s'agit généralement d'un histogramme du nombre d'occurrences de motifs visuels spécifiques dans une image donnée. Les motifs visuels sont appelés des mots qui sont pré-construits dans un livre (codebook) à l'aide de techniques de regroupement (Clustering). Malgré leur simplicité, le sac de mots visuels a bien été appliqué avec succès à divers cas difficiles de vision par ordinateur, y compris des études récentes visant à explorer leur applicabilité à la reconnaissance automatique des activités humaines. Cependant, l'indexation des activités humaines est encore à ses débuts en raison de la complexité des tâches à accomplir. Dans l'article

de *Niebles et al.* [162], les auteurs ont présenté une approche pour la classification non supervisée des actions humaines en différentes catégories à partir de séquences vidéo. La base de leur méthode est l'extraction d'une collection des mots spatio-temporels via l'utilisation de modèles de sujets latents (latent topic models).

3.6.4 Divertissement

Ces dernières années, l'industrie du jeu a attiré un nombre croissant de personnes. Une nouvelle génération de jeux basés sur le jeu complet du corps comme les jeux de sport a augmenté l'attrait du jeu pour les membres de la famille de tout âge. Pour permettre une perception précise des actions humaines, ces jeux utilisent des capteurs RGB-D économiques (Par exemple, Kinect [194]), qui fournissent des données supplémentaires sur le canal de profondeur [229, 232]. Ces données de profondeur codent des informations structurelles riches de la scène entière et facilitent la tâche de reconnaissance d'action en simplifiant les variations de mouvement intra-classe et en réduisant le bruit de fond [120, 122].

3.6.5 Véhicule de conduite autonome

Les algorithmes de prédiction des actions [121] pourraient être l'un des éléments potentiels et peut-être les plus importants d'un véhicule autonome. Les algorithmes de prédiction d'action peuvent prédire l'intention d'une personne [139] dans un court laps de temps. Dans une situation d'urgence, un véhicule équipé d'un algorithme de prédiction d'action peut prédire l'action ou la trajectoire future d'un piéton dans quelques secondes qui suivent, ce qui pourrait être essentiel pour éviter une collision. Ces méthodes sont basées sur l'analyse les caractéristiques de mouvement du corps humain à un stade précoce d'une action à l'aide de points d'intérêt ou de réseaux neuronaux convolutifs [124]. Les algorithmes de prédiction d'action peuvent comprendre les actions possibles en analysant l'évolution des actions sans avoir observé l'exécution complète des actions.

3.7 Bases de données

Les bases de données ou le jeu de données publiques constituent un critère commun pour mesurer et comparer la validité des approches proposées. La plupart des premiers jeux de données sont construits avec une seule caméra contenant une douzaine d'actions simples pour un nombre limité de personnes. L'enregistrement se fait habituellement dans un environnement contrôlé avec des réglages simples. Les annotations se font par les créateurs de base de données de manière manuelle. Récemment, les nouveaux jeux de données ont été créés en téléchargeant des clips vidéo et des films en ligne. Il existe plusieurs classifications des bases de données de la reconnaissance des activités humaines [1, 38]. En outre, les bases sont regroupées soit selon la finalité de l'application visée ou selon la manière de la construction de ces bases. La table 3.1 montre des exemples de base de données publique pour la reconnaissance des activités humaines regroupées en quatre catégories : les bases de données réalistes, irréalistes, analyse des interactions et les bases avec Multiview.

La table 3.2 montre aussi une liste des bases de données publique divisées en deux catégories : contrôlée ou non contrôlée. Dans notre cas, on s'intéresse seulement aux bases de deux premières catégories : contrôlées ou irréalistes et non contrôlées ou réalistes. Dans les sections suivantes, nous avons sélectionné deux bases les plus célèbres du domaine appartenant aux deux catégories choisies.

3.7.1 Bases de données irréalistes

Les bases de données irréalistes sont appelées parfois les bases de données contrôlées, cela est dû au fait qu'elles reposent sur la collection des vidéos avec des fortes contraintes d'acquisitions telles que la caméra fixe, le font immobile, un seul sujet, etc.

3.7.1.1 Base de données KTH

KTH [190] est une base de données contrôlées, car elle possède de fortes contraintes d'acquisitions comme l'homogénéité de l'arrière-plan, caméra fixe, etc.

Table 3.1 – Les bases de données par catégories de la reconnaissance des activités humaines

Type de problème	Base de données	Année
Irréalistes	Weizmann	2005
	KTH	2004
Réalistes	CAVIAR	2004
	ETISEO	2005
	CASIA Action	2007
	MSR Action	2009
	UT Tower	2010
	HOLLYWOOD	2009
	UCF SPORT	2008
	UCF YouTube	2009
	UCF 101	2013
	Olymbic	2010
	HMDB51	2011
Analyse des intéraction	BEHAVE	2004
	TV Human Interaction	2010
	UT Interaction	2010
Analyse de multiview	IXMAS	2006
	i3DPost Multi-view	2009
	MuHAVI	2010
	VideoWeb	2010

Elle permet d'étudier, d'analyser et de reconnaître des actions humaines, cette base est considérée comme la plus ancienne dans ce domaine. En outre, elle devenu un standard et nous considérons comme étant la première étape de la validation des approches proposées par les nouveaux chercheurs du domaine de la reconnaissance des actions humaines à partir des vidéos.

La base de données vidéo actuelle contient six types d'actions humaines (*walking, jogging, running, boxing, hand waving and hand clapping*) réalisées plusieurs fois par 25 sujets dans quatre scénarios différents : en plein air *s1*, en plein air avec variation d'échelle *s2*, en plein air avec différents vêtements *s3* et en intérieur *s4* (Figure 3.15). Actuellement, la base de données contient 2391 séquences, toutes les séquences ont été prises sur des fonds homogènes avec une caméra statique à 25 images/seconde ; Les séquences ont été échantillonnées à une résolution spatiale de 160 x 120 pixels et leur durée moyenne est de quatre secondes.

Table 3.2 – Les bases de données vidéo populaires utilisés dans la recherche des activités humaines. (Source avec modification [123])

Base de données	Année	Vidéos	Actions	Sujets	Env.
KTH	2004	599	6	25	Contrôlé
Weizmann	2005	90	10	9	Contrôlé
INRIA XMAS	2006	390	13	10(3 times)	Contrôlé
IXMAS	2006	1,148	11	-	Contrôlé
UCF Sports	2008	150	10	-	N.Contrôlé
Hollywood	2008	-	8	-	N.Contrôlé
Hollywood2	2009	3,669	12	10	N.Contrôlé
UCF 11	2009	1,100+	11	-	N.Contrôlé
MSR-I	2009	63	3	10	Contrôlé
MSR-II	2010	54	3	-	Foule
TV-I	2010	300	4	-	N.Contrôlé
MSR-A	2010	567	20	1	Contrôlé
Olympic	2010	783	16	-	N.Contrôlé
HMDB51	2011	7,000	51	-	N.Contrôlé
CAD-60	2011	60	12	4	Contrôlé
BIT-I	2012	400	8	50	Contrôlé
LIRIS	2012	828	10	-	Contrôlé
MSRDA	2012	320	16	10	Contrôlé
UCF50	2012	50	50	-	N.Contrôlé
UCF101	2012	13,320	101	-	N.Contrôlé
MSR-G	2012	336	12	1	Contrôlé
UTKinect-A	2012	10	10	-	Contrôlé
ASLAN	2012	3,698	432	-	N.Contrôlé
MSRAP	2013	360	6 pairs	10	Contrôlé
CAD-120	2013	120	10	4	Contrôlé
Sports-1M	2014	1,133,158	487	-	N.Contrôlé
3D Online	2014	567	20	-	N.Contrôlé
FCVID	2015	91,233	239	-	N.Contrôlé
ActivityNet	2015	28,000	203	-	N.Contrôlé
YouTube-8M	2016	8,000,000	4,716	-	N.Contrôlé
Charades	2016	9,848	157	-	Contrôlé
NEU-UB	2017	600	6	20	Contrôlé
Kinetics	2017	500,000	600	-	N.Contrôlé
AVA	2017	57,600	80	-	N.Contrôlé
20BN-Something	2017	108,499	174	-	N.Contrôlé
SLAC	2017	520,000	200	-	N.Contrôlé
Moments in Time	2017	1,000,000	339	-	N.Contrôlé

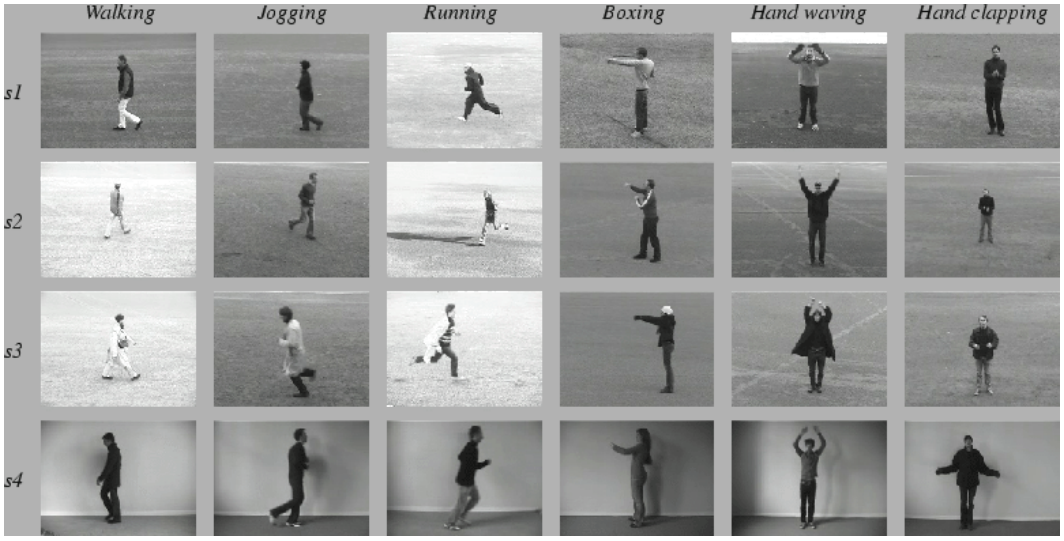


Figure 3.15 – Illustration de la base de données KTH [190].

3.7.1.2 Base de données Weizmann

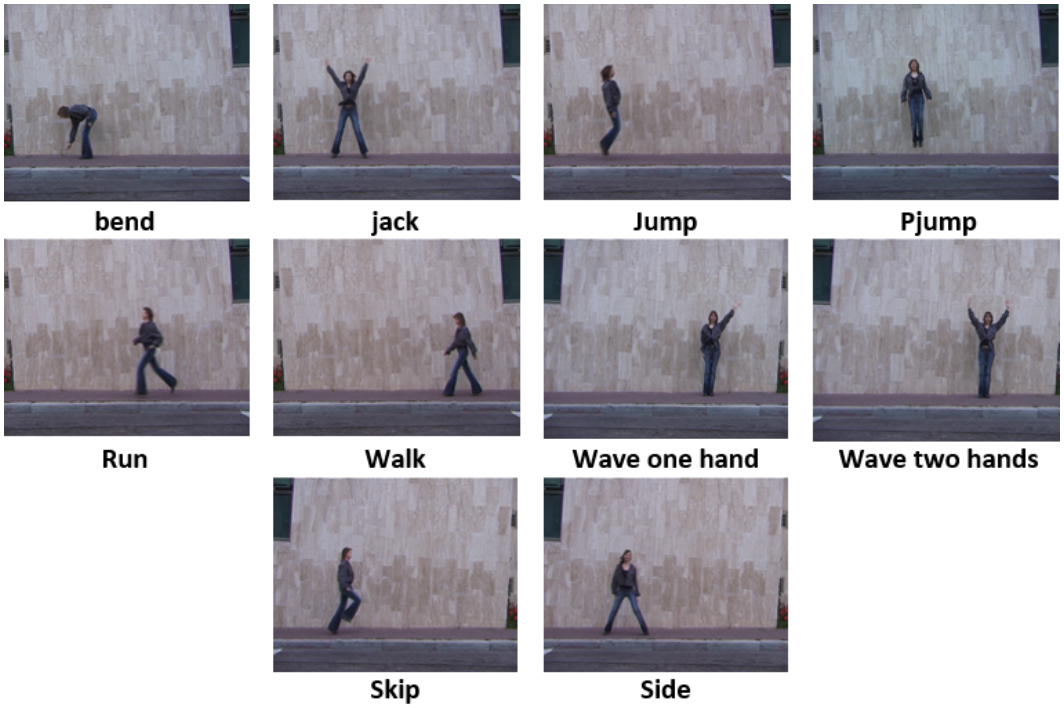


Figure 3.16 – Illustration de la base de données Weizmann [20, 86].

La base de données Weizmann [20, 86] contient une collection de 90 séquences vidéo en basse résolution (180x144, 50 images /seconde) montrant neuf personnes différentes, chacune effectuant 10 actions naturelles : *"run", "walk", "skip",*

"jumping-jack" (ou simplement "jack"), "jump-forward-on-two-legs"(ou "jump"), "jump-in-place-on-two-legs" (ou "pjump"), "gallop-sideways" (ou "side"), "wave-two-hands" (ou "wave2"), "wave-one-hand" (ou "wave1"), ou "bend". Cette base de données est similaire à celle de KTH dans le contexte des contraintes de l'acquisition des vidéos. La figure 3.16 montre l'ensemble des classes d'actions de cette base de données.

3.7.2 Base de données réalistes

Les bases de données non contrôlés ou réalistes sont collectées à partir des sites Web de stockage des vidéos, avec de faibles contraintes d'acquisitions (caméra non fixe, occultations partielles, changement de point de vue, la résolution, etc.). Ces vidéos montrent des actions humaines dans des contextes quotidiens. Nous présentons ci-dessous deux bases de données les plus célèbres et les plus utilisées par la communauté scientifique de la vision par ordinateur.

3.7.2.1 Base de données UCF101

UCF101 [197] est un jeu de données créé pour la validation des modèles proposés pour la reconnaissance des activités humaines réelles, recueillies à partir de YouTube, ayant 101 catégories d'activités quotidiennes (Figure 3.17). Ce jeu de données est une extension de l'ensemble de données UCF50 qui a 50 catégories d'action , avec 13320 vidéos de 101 catégories d'action, UCF101 offre la plus grande diversité en termes d'actions et avec la présence de grandes variations dans le mouvement de la caméra, l'apparence et la pose des objets, l'échelle des objets, le point de vue, le fond encombré, les conditions d'éclairage, etc, c'est le jeu de données le plus difficile. Les vidéos de 101 catégories sont regroupées en 25 groupes, où chaque groupe peut se composer de 4 à 7 vidéos de chaque action, les vidéos d'un même groupe peuvent partager certaines caractéristiques communes, comme un arrière-plan similaire, un point de vue similaire, etc.

Les catégories d'actions peuvent être divisées en cinq types :

Figure 3.17 – Illustration de la base de données UCF101 [197].

— **Interaction homme-objet**

— **Corps en mouvement seulement**

— **Interaction homme-homme**

— **Pratique d'instruments de musique**

— **Sport.**

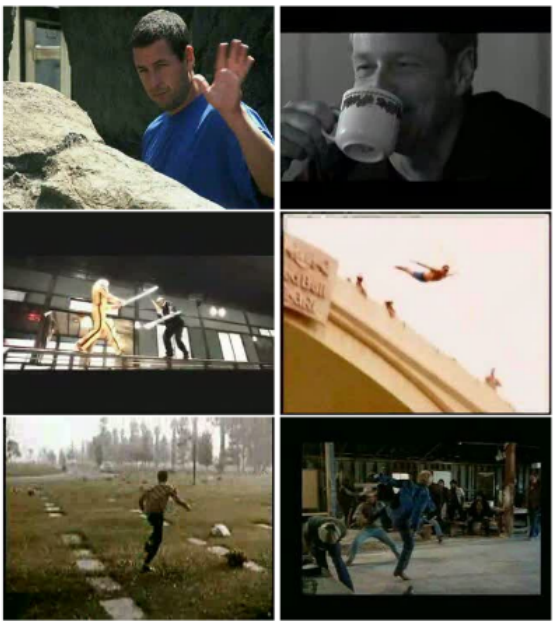


Figure 3.18 – Illustration de la base de données HMDB51 [127].

3.7.2.2 Base de données HMDB51

HMDB [127] est collectée à partir de diverses sources, principalement des films, et d'une petite partie des bases de données publiques telles que les archives de Prelinger, YouTube et Google vidéos. L'ensemble de données contient 6 849 clips répartis en 51 catégories d'actions, chacune contenant au minimum 101 clips (Figure 3.18). Les catégories d'actions peuvent être regroupées en cinq types :

- **Actions faciales générales** : sourire, rire, ...
- **Actions faciales avec manipulation des objets** : fumer, manger, ...
- **Mouvements corporels généraux** : frappe des mains, monter, ...
- **Mouvements corporels avec interaction des objets** : brosser les cheveux, dribbler, jouer au golf, ...
- **Mouvements corporels pour l'interaction humaine** : coup de pied à quelqu'un, coup de poing, ...

3.7.3 Base de données et les modèles de la reconnaissance

Dans cette partie, nous allons résumer le plus important modèle et architectures de la reconnaissance des activités humaines. Généralement, on peut noter que les architectures proposées pour la reconnaissance des actions humaines peuvent être classées en trois grandes catégories principales : les architectures basées sur la représentation des actions humaines et la classification à base de ses représentations, les architectures à base de Deep Learning et enfin les architectures hybrides.

À cette fin, nous avons mené une recherche bibliographique pour identifier les modèles les plus importants utilisés dans ce domaine. Par ailleurs, nous avons limité notre recherche sur les architectures appliquées sur deux bases de données réelles UCF101 et HMDB51 et deux bases de données irréalistes KTH et Weizmann.

3.7.3.1 Modèles basés représentation

La table 3.3 montre quelques modèles basés sur les techniques de la représentation des actions humaines (locale, globale, modèle, pose, ...) comme une étape nécessaire pour la classification des actions humaines.

Table 3.3 – Les modèles à base de représentation.

Modèles	Réf	Base de données			
		KTH	Weiz	UCF101	HMDB51
Dense Traj (Traj + HOG+HOF+MBH)	[211]	94.2%		88.2%(UCF-Sport)	
Motion Interchange Patterns	[118]	93%		68.5%(UCF-50)	29.2%
MBH + SIFT + Sqrt + L2 Normalization	[168]			90%(UCF-50)	54.8%
Improve Traj (Without Human Detector)	[212]			90.5%(UCF-50)	55.9%
Improve Traj (With Human Detector)	[212]			91.2%(UCF-50)	57.2%
Traj + HoG + HoF + MBH + DCS on w-flow	[103]				52.1%
Stacked FVs + FV	[172]				66.8%
Hybrid-BoW	[171]			87.9%	61.1%
MPEG-Flow video descriptor	[108]			85.6%(UCF-50)	46.3%
SDT tree ATEP	[75]				41.3%
TrajShape+TrajMF	[106]			78.5%	48.4%
Rank Pooling	[72]				63.7%
Multi-Skip Feat. Stacking	[133]			89.1%	65.1%
Multi-channel correlation filters	[115]		97.80%	82.6%(UCF-Sport)	
Interest points + SIFT filters	[157].	97.89%	96.66%		
Binary motion descriptor	[66].	96%	95.81%		
Shape, motion and texture features	[178].	94.91%	94.44%		
Pose primitive	[201].	70%	94.40%		
Sequence alignment and shape context	[7].		92.22%		
Hough Transform-Based Voting Framework	[233].	93.5%	97.8%	86.6%	
Learning Mid-level Motion Features	[67].	90.50%	90.50%		
Spatial-Temporal	[162]	83.33%	90.00%		

3.7.3.2 Modèles basés Deep Learning

La table 3.4 montre quelques modèles à base de Deep Learning pour la reconnaissance des actions humaines.

Table 3.4 – Les modèles à base de Deep Learning.

Modèles	Réf	Base de données			
		KTH	Weiz	UCF101	HMDB51
Binary Motion Image and Deep Learning with 5 classes	[63]		100% (5 classes)		
Two-stream (CNN-M-2048)	[195]			88.0%	59.4%
Transfer Learning on Sports 1M	[112]			65.4%	
Factorized Spatio Temporal Conv. Nets	[200]			88.1%	59.1%
Two-Stream (ClarifaiNet)	[218]			88.0%	
Two-Stream (GoogLeNet)	[218]			89. 3%	
Two-Stream (VGG-16)	[218]			91.4%	
Conv Pooling (Image + Opt Flow)	[238]			88.2%	
LSTM (Image + Opt Flow)	[238]			88.6%	
Adaptive Multi-Stream Fusion	[228]			92.6%	
C3D on SVM	[204]			85.2%	
ImageNet pretrain + tuple verification	[155]			30.6%	29.9%
RBG + Opt Flow Networks	[221]			92.4%	62%
End to End Rank-pooling	[73]			87% (Sport)	
Hierarchical Rank-pooling (CNN Features)	[71]			91.4%	66.9%
Two Stream Fusion (VGG-16)	[69]			92.5%	65.4%
LTC (flow+RGB)	[208]			92.7%	67.2%

3.7.3.3 Modèles hybrides

La table 3.5 montre quelque exemples des modèles hybrides (Représentation + Deep Learning).

Table 3.5 – Les modèles hybrides.

Modèles	Réf	Base de données			
		KTH	Weiz	UCF101	HMDB51
TDD + Human detector	[217]			91.5%	65.9%
TDD	[217]			90.3%	63.2%
Hierarchical RP on CNN + Rank Pooling	[71]			90.7%	65.0%
VLAD	[141]			84.7%	
VLAD + Human detector	[141]			92.2%	
LTC (flow+RGB) + Human detector	[208]			92.7%	67.2%
Two Stream Fusion (VGG-16) + Human detector	[69]			93.5%	69.2%
Hybrid fusion + Deepnets	[53]			92.5%	70.4%
C3D + Human detector on SVM	[204]			90.4%	

3.8 Conclusion

Dans ce chapitre, nous avons présenté l'architecture générale d'un système de la reconnaissance des activités humaines et un état de l'art sur ce domaine. Nous avons montré que cette architecture est composée de trois niveaux de base et chaque niveau est constitué d'un ensemble d'étapes. Nous avons concentré sur le premier niveau ou l'architecture de base. Cette dernière est composée de trois composantes : la détection et le suivi de l'objet, l'extraction et la représentation des caractéristiques et enfin l'étape de classification. Par la suite, nous avons expliqué longuement les différents types d'approches qui sont adoptées par les chercheurs du domaine. À la fin, nous avons présenté quelques catégories de bases de données utilisées pour la validation des approches dans le domaine de reconnaissance des activités humaines.

Dans les chapitres suivants, nous présenterons notre modèle de la reconnaissance des actions humaines à partir des séquences vidéo. Ce modèle est basé sur une méthode hybride d'extraction des caractéristiques locales et globales à partir des séquences vidéo en se basant sur l'estimation de flux optique. En fonction de ses caractéristiques, le descripteur est créé pour la classification des actions. Afin d'exploiter en profondeur le modèle proposé, nous avons proposé une méthode de décomposition des activités complexes en des actions simples. Le dernier chapitre de la thèse concerne le processus de validation du système proposé.

4 Descripteur de mouvement proposé

4.1 Introduction

Dans ce chapitre et les chapitres suivants, nous discuterons en détail notre modèle proposé pour la reconnaissance automatique des actions humaines. Ce chapitre se base principalement sur un ensemble de publications scientifiques présenté lors de conférences scientifiques spécialisées [131, 130], deux publications dans des revues scientifiques reconnues [132, 82] ainsi que une publication d'un chapitre dans un livre spécialisé dans le domaine de la surveillance visuelle intelligente [129]. Notre modèle s'articule sur trois grandes étapes :

— Extraction des caractéristiques de mouvement enregistrés dans une séquence vidéo en se basant seulement sur l'estimation du flux optique ;

— Représentation de ses caractéristiques sous forme d'un descripteur.

— Reconnaissance des actions humaines par l'application d'un processus de classification.

Ce chapitre est consacré aux deux premiers points et le dernier point sera présenté dans un chapitre à part. La suggestion d'une approche particulière nécessite à un moment donné d'examiner la fiabilité et la performance de cette proposition.

C'est pour ça que nous avons proposé un autre chapitre concerne l'analyse de la performance de descripteur proposé afin de compléter l'étude de ce sujet sous tous ses aspects.

En ce qui concerne ce chapitre, il reflète nos contributions les plus importantes, dans lesquelles nous clarifions les principes fondamentaux sur lesquels est basé le descripteur du mouvement proposé . Avant de discuter de cela, une étape importante doit d'abord être franchie : il faut détecter les mouvements des humains dans les séquences vidéo et extraire leurs caractéristiques de base pour les utiliser lors de l'étape de construction du descripteur. Cette étape de détection et l'extraction des caractéristiques des mouvements sont exclusivement dépendantes d'un outil puissant celui de flux optique. Pour cela notre descripteur est nommé "*Flow Motion Descriptor*" ou "*descripteur de mouvement basé flux*".

4.2 Approche proposée

En raison du rôle indiscutable des systèmes automatiques de la reconnaissance des activités humaines dans plusieurs domaines d'application, nous proposons dans cette thèse un système pour l'analyse et la classification des actions humaines. Ce système se compose de trois éléments principaux l'hypothèse suivante doit être vérifiée : " *la vidéo ne doit contenir que des mouvements liés à une action humaine*". Dans un premier temps, le flux optique est estimé au moyen d'un ensemble consécutif de trames. Ensuite, le vecteur de caractéristiques est construit sous forme d'histogramme à partir des descripteurs de mouvement pour les trames considérées. La classification des actions est effectuée à l'aide de classifieurs basés sur un sous-ensemble de caractéristiques extrait pendant la phase de sélection des caractéristiques.

4.2.1 Principe

Nous avons montré dans le chapitre précédent (Chapitre III), un aperçu général d'un système de la reconnaissance des actions humaines basées sur la vision

par ordinateur. Pour notre part, nous avons proposé le digramme 4.1 qui précise les étapes de manière détaillées de ce processus. Cette approche est composée de trois phases, la détection et le suivi de personne, l'extraction et la représentation des caractéristiques et enfin l'étape de la reconnaissance des actions par l'application d'un ensemble des classifieurs. Pour la détection des personnes, les histogrammes des gradients orientés ou (***HOG :Histograms of Oriented Gradients***) basés sur le détecteur de piétons de *Dalal et al.* [50] peuvent être utilisés. Ce détecteur est le plus performant en matière de détection des piétons de sorte que l'ensemble du corps est détecté. Afin d'augmenter le rappel de la détection de personnes dans des conditions difficiles, une approche simple de suivi des personnes est utilisée. La zone de délimitation (bounding box) de chaque piéton détecté se propage ensuite à la trame suivante [26]. Récemment, de nombreuses autre techniques ont émergé pour résoudre ce problème, on peut les trouver bien détaillées dans les articles [161, 163]

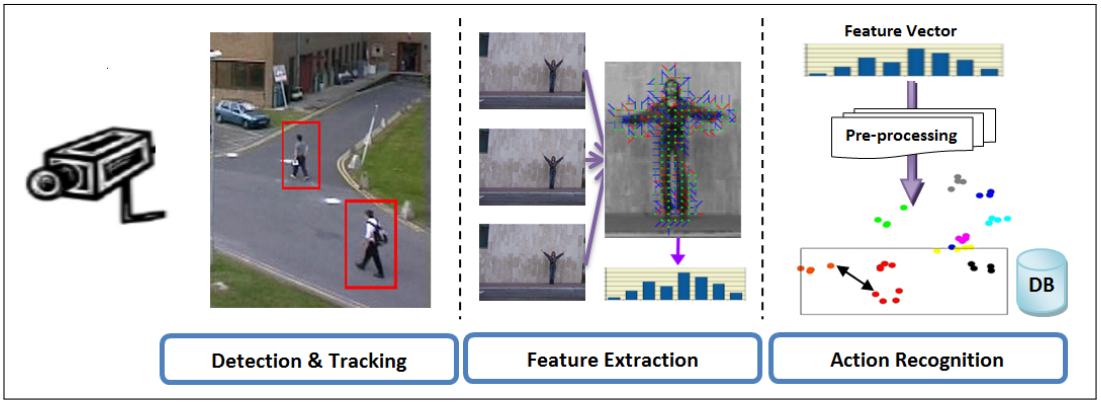


FIGURE 4.1 – Vue générale du système proposé pour la reconnaissance des actions humaines.

Premièrement le corps humain est sélectionné préalablement de l'image, puis vient l'étape d'extraction des caractéristiques des mouvements via l'estimation de flux optique entre les trames successives. Par la suite, un descripteur de mouvement est créé par la construction des histogrammes de mouvement locaux et globaux. La sélection des caractéristiques est un processus utilisé afin d'améliorer la performance de la classification. Enfin, plusieurs classifieurs utilisent les descripteurs calculés précédemment pour déterminer la classe de l'action de chaque séquence vidéo, l'étape de l'analyse est la dernière étape pour identifier les points forts et les faiblesses de notre système proposé.

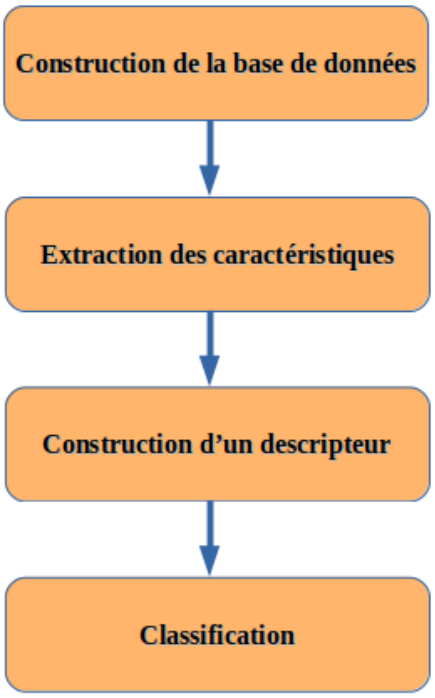


Figure 4.2 – Architecture détaillée du système proposé pour la reconnaissance des actions humaines

Le diagramme de la figure 4.2 illustre le processus que nous avons entrepris et les étapes à suivre dans notre système afin de reconnaître des actions humaines. La première étape est une étape de prétraitement permettant la construction des bases de données des actions simples à partir des bases de données standard des activités humaines. Cette étape de prétraitement est bien détaillée dans la section (Section 4.3). La section (Section 4.4) est consacrée à la notion de l'estimation de mouvement et le flux optique. L'extraction des caractéristiques est l'étape suivante dans notre diagramme indiquée dans la section (Section 4.5). Dans le même chapitre et dans la section (Section 4.6), nous présentons le processus de la construction du descripteur. Dans le chapitre suivant (Chapitre V) , nous détaillerons la phase de classification.

Avant de détailler ces étapes de ce système, nous aimerions poser la question suivante : ce système est-il capable de connaître les actions complexes ? Nous suggérons de répondre à cette question dans la section suivante et dans le chapitre suivant.

4.2.2 Au-delà des actions simples

Récemment, la communauté des chercheurs de la vision par ordinateur est orientée vers un processus d'exploitation en profondeur des actions humaines de base en vue de la reconnaissance automatisée des activités humaines dans des scènes complexes. On peut citer les publications suivantes :

— **Yi *et al*** [235] ont décrit l'évolution du mouvement humain pour classifier des actions complexes. L'information structurelle temporelle est dérivée à partir d'une trame clé sélectionnée où une représentation vidéo hiérarchique est proposée en fonction de la trajectoire pour le codage des clips vidéo aux différents niveaux.

— **Feng *et al.*** [70] ont exploité l'utilisation d'un processus minier pour les modèles spatiaux-temporels afin de construire une méthode de dénoiement des mouvements humains fondée sur des données. La détection des actions de base et des modèles de mouvement est effectuée à l'aide d'une méthode d'apprentissage par dictionnaire où de multiples mots-clés compacts et représentatifs de mouvement qui sont appris à partir des données d'apprentissage.

— **Alfaro *et al.*** [5] ont proposé une méthode pour réduire une vidéo en un ensemble de séquences clés représentant des actes atomiques significatifs de chaque classe d'action.

— **Zhu *et al.*** [240] ont proposé une approche appelée key volume mining deep framework. Le framework est basé sur des volumes miniers clés ou rudimentaires pour chaque classe d'action humaine.

— **Ladjailia *et al.*** [132] ont proposé une méthode permettant la décomposition des activités complexe en une liste des actions simple ou des mots en se basant sur des actions clés prédéfinies. D'autre part, la définition de l'activité comme étant un ensemble ou une séquence d'actions permet d'identifier une activité complexe à travers ses actions élémentaires et le modèle de cette activité.

4.3 Construction de la base de données des actions

La construction d'une base de données des actions simples ou élémentaires est une étape de prétraitement. Cette étape vise à créer une base de données des actions simples ou élémentaires à partir des bases de données des activités complexes tels que UCF101, HMDB51,... On peut aussi extraire les actions simples à partir des bases de données telles que Weizmann, KTH,... qui contiennent des activités simples ou, en d'autres termes, des actions répétitives simples. Notre mission consiste à extraire un cycle de mouvement (une action simple) à partir de la vidéo d'une activité ou plusieurs activités. Cette tâche est effectuée manuellement. En revanche, sur la base d'expériences empiriques, une action simple peut être suffisamment représentée en utilisant un ensemble de 15 trames consécutives pour les vidéos capturées avec une fréquence de 25 trames /secondes.

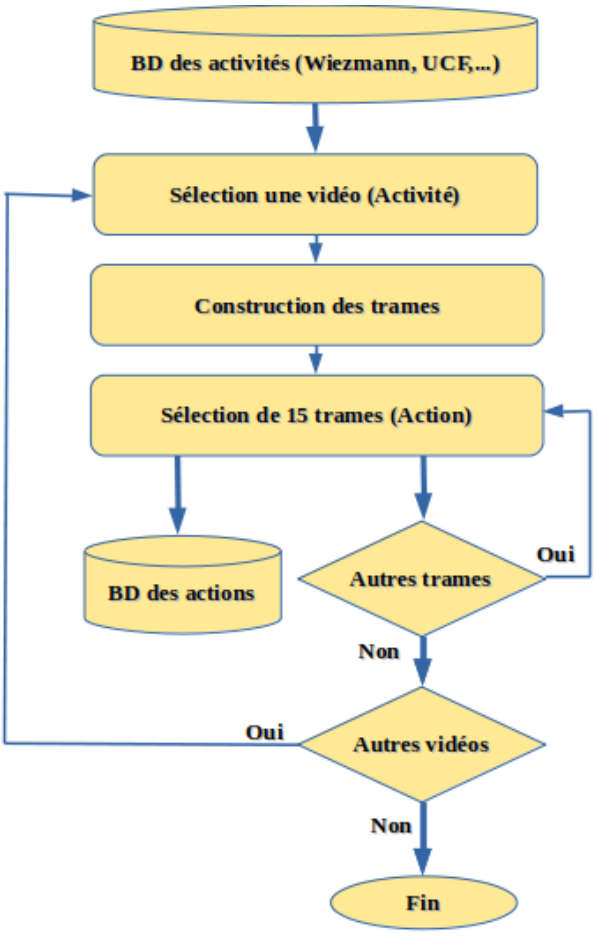


Figure 4.3 – Construction de la base de données des actions

Le diagramme 4.3 résume les étapes de la construction des bases de données des actions simples à partir des bases de données des activités.

4.4 Estimation de mouvement (Motion estimation)

L'estimation du mouvement est un procédé qui consiste à étudier le déplacement des objets dans une séquence vidéo, en cherchant la corrélation entre deux images successives afin de prédire le changement de position du contenu. Le mouvement est un réel problème en vidéo puisqu'il décrit un contexte en trois dimensions alors que les images sont une projection de scènes 3D dans un plan en 2D [202]. Le plus souvent, le terme estimation de mouvement et le terme flux optique sont utilisés de façon interchangeable. En fait, le flux optique consiste à estimer le mouvement, c'est-à-dire à calculer la vitesse du mouvement sur l'image 2D.

Notre approche proposée est basée sur l'extraction des caractéristiques des mouvements à partir des séquences vidéo à l'aide de l'estimation de flux optique entre les trames successives. Évidemment, le premier point à aborder est la notion de flux optique lui-même.

4.4.1 Notion de flux optique

Le concept de ***flux optique*** ou "***flot optique***" a été introduit dans plusieurs domaines :

— En *psychologue*, l'américain *James J. Gibson* est inventé ce concept dans les années 1940 pour décrire la perception visuelle des mouvements chez les animaux [79].

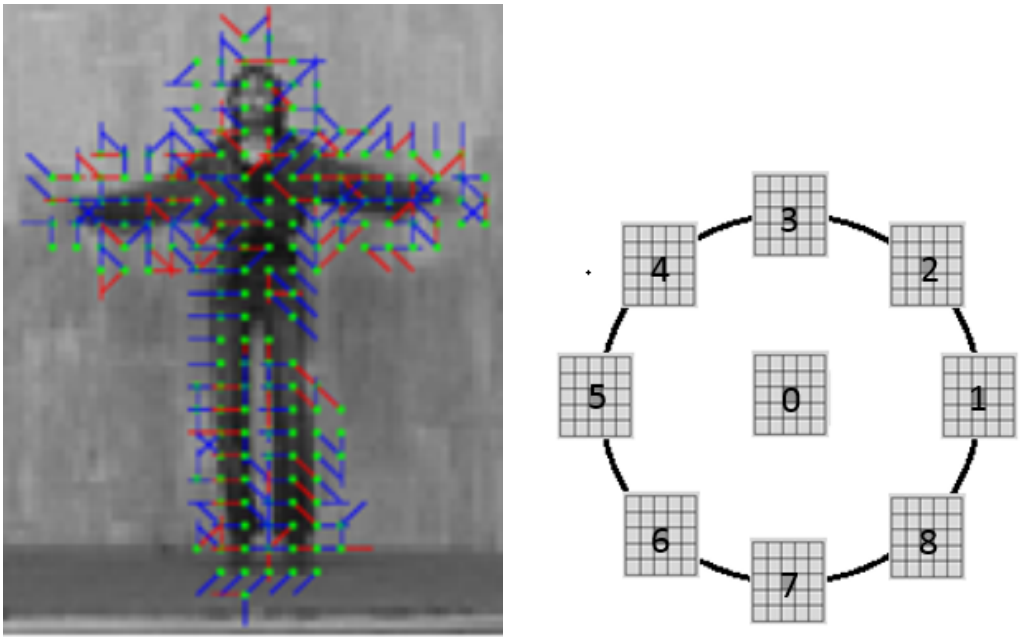
— Dans le domaine de *la vision par ordinateur*, Il existe plusieurs définitions de ce concept, par exemple :

* Selon *Burton et Warren* [30, 223], le flux optique est *le mouvement apparent des objets, surfaces et contours d'une scène visuelle, causé par le mouvement relatif entre un observateur et la scène.*
* Selon *Horn et al.* [95, 94], le flux optique est *la projection 2D des mouvements des objets 3D en mouvements.* On remarque que cette définition a négligé le rôle de l'observateur.
* Les deux définitions ci-dessus peuvent être combinées comme suit : *"Le flux optique est le mouvement apparent des objets entre deux images consécutives causé par le mouvement de l'objet ou de la caméra. Il s'agit d'un champ vectoriel 2D où chaque vecteur est un vecteur de déplacement montrant le mouvement des points de la première image à la seconde"*. En bref *"Le flux optique est une projection d'un mouvement du monde réel sur une image plane"*.

4.4.2 Algorithmes de l'estimation de mouvement

Plusieurs algorithmes ont été proposés pour estimer le mouvement apparent et l'étudier théoriquement et expérimentalement. Les algorithmes les plus connus dans ce domaine sont ceux appartenant à un ensemble des méthodes dites différentielles ou celles classées comme des algorithmes du type Matching ou (Appariement|correspondance). La figure 4.4 montre un exemple de chaque type.

Les méthodes différentielles estiment la vitesse de l'image à partir de la variation spatiale et temporelle de la luminosité de l'image. En revanche, ***Les méthodes du type Matching recherchent les déplacements qui mettent en correspondance les caractéristiques de luminosité de l'image.*** Ces deux méthodes ont des applications dans différents domaines. Les méthodes différentielles sont les meilleures lorsque les déplacements dans l'image sont petits (2 pixels) et les méthodes de Matching fonctionnent bien pour les déplacements modérés mais ne manipulent pas les mouvements au niveau pixels [144]. Les deux types d'algorithmes de flux optique peuvent utiliser des contraintes locales ou globales.



(a) Flux Optique : Estimation de la vitesse de la variation spatiale et temporelle.

(b) Bloc matching : Estimation de similarité entre le bloc de l'image d'origine et leurs voisins dans les images suivantes.

Figure 4.4 – Techniques de l'estimation de flux optique

Dans notre étude, nous avons choisi le flux optique comme une méthode différentielle pour estimer les mouvements de l'être humain dans une séquence vidéo. En pratique, le calcule de ce flux doit être estimé à partir des variations spatio-temporelles de l'intensité de l'image. Cela entraîne immédiatement des difficultés, car il y a des cas où *un objet en mouvement ne produit aucune variation dans l'image*, par exemple une sphère tournante uniforme [94]. De plus, une partie de la variation de l'intensité de la scène ne peut pas être due au mouvement, mais à des *variations de l'éclairage*. Néanmoins, la première hypothèse qui concerne la conservation de l'intensité des points de la scène est adoptée dans notre travail, ainsi que toutes les variations dans la séquence d'images sont dues au mouvement.

4.4.2.1 Hypothèses

Les méthodes à base de flux optique tentent de calculer le mouvement entre deux trames d'image prises aux instants t et $t + \Delta t$ pour chaque pixel. Ces méthodes sont *des méthodes différentielles* car elles sont basées sur des approximations locales en série de Taylor du signal d'image ; c'est-à-dire qu'ils utilisent des dérivées partielles par rapport aux coordonnées spatiales et temporelles.

Les méthodes différentielles sont basées sur les contraintes de ***la conservation des couleurs ou de l'intensité des pixels*** :

Definition 4.1. La constance des couleurs (Color constancy)

La constance des couleurs est la conservation de la couleur d'un pixel dans une trame de référence à son nouvel emplacement dans une trame cible.

Dans le cas particulier où l'image est au niveau du, le principe ci-dessus devient :

Definition 4.2. Conservation de l'Illumination ou conservation d'intensité(brightness constancy)

Conservation de l'Illumination est la conservation de la luminosité d'un point dans une trame de référence à son nouvel emplacement dans une trame cible.

4.4.2.2 Estimation

Supposant qu'un pixel à l'emplacement (x, y, t) et avec l'intensité $I(x, y, t)$ aura été déplacé de Δx dans la X-direction et Δy dans Y-direction dans le temps Δt vers un pixel de la trame suivante et avec l'intensité $I(x + \Delta x, y + \Delta y, t + \Delta t)$. La luminosité du pixel est supposée rester constante selon les contraintes ci-dessus. En mathématiques, on peut écrire cette conservation de l'intensité sous la forme :

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{4.1}$$

Si on applique le développement de Taylor du côté droit de l'équation 4.1, nous obtenons,

$$I(x+\Delta x, y+\Delta y, t+\Delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \epsilon \qquad (4.2)$$

le ϵ contient des termes de second ordre les plus élevés de développement limité de Taylor. Après la soustraction de $I(x,y,t)$ des deux côtés de l'équation précédente, on a trouvé :

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \epsilon = 0 \qquad (4.3)$$

Après la divisions les deux côtés par Δt, on a obtenu :

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t} + \epsilon(\Delta t) = 0 \qquad (4.4)$$

Sachant que $\epsilon(\Delta t)$ est un terme qui dépend de Δt et

$$\lim_{\Delta t \to \mathbf{0}} \epsilon(\Delta t) = 0$$

Donc, l'équation 4.4 devient :

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} = 0 \qquad (4.5)$$

Qui se traduit par :

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \qquad (4.6)$$

Sachant que $V_x = \frac{\Delta x}{\Delta t}$, $V_y = \frac{\Delta y}{\Delta t}$ sont la vitesse ou le flux optique de point $I(x, y, t)$ selon la X-direction et Y-direction respectivement. Il s'agit d'une équation à deux inconnues qui ne peut pas être résolue comme telle. C'est un problème du type ***malle posée*** (Aperture Problem) des algorithmes de flux optique. Pour trouver le flux optique, on a besoin des contraintes supplémentaires pour résoudre ce problème. Toutes les méthodes de flux optique introduisent des conditions supplémentaires pour estimer le flux réel.

4.4.2.3 Méthodes de détermination

Dans la littérature il existe plusieurs méthodes pour la résolution de l'équation 4.6. Chaque méthode a ses avantages et ses inconvénients, nous citons ci-dessous quelques exemples de ces méthodes :

— **Méthode de corrélation de phase [184, 78] :** La corrélation de phase est une approche permettant d'estimer le décalage de translation relative entre deux images similaires (corrélation d'images numériques) ou d'autres ensembles de données. Il est couramment utilisé dans l'enregistrement d'images et repose sur une représentation des données dans le domaine des fréquences, généralement calculées par la transformée de Fourier rapide. La technique s'applique particulièrement à un sous-ensemble de techniques de corrélation croisée qui isolent l'information de phase de la représentation spatiale de Fourier du corrélogramme croisé.

— **Méthodes basées sur les blocs [117] :** Leur principe consiste à minimiser la somme des différences au carré ou la somme des différences absolues, ou maximiser la corrélation croisée normalisée.

— **Méthodes d'optimisation discrètes** L'espace de recherche est quantifié, puis l'appariement d'images est traité par l'attribution d'étiquettes à chaque pixel, de sorte que la déformation correspondante minimise la distance entre la source et l'image cible [81]. La solution optimale est souvent récupérée par des algorithmes basés sur le théorème de coupe minimale Max-flow, la programmation linéaire ou les méthodes de propagation des croyances.

— **Méthodes différentielles :** Elles sont basées sur les dérivées partielles du signal d'image et/ou du champ de flux recherché et des dérivées partielles d'ordre supérieur, parmi les méthodes de ce type, on peut citer :

* **Méthode Lucas-Kanade :** Cette méthode est proposée par *Lucas et al.* [149] suppose que le déplacement du contenu de l'image entre deux instants proches est faible et approximativement constant dans un voisinage du point p considéré. On peut donc supposer que l'équation du flux optique est valable pour tous les pixels d'une fenêtre centrée sur p.
* **Méthode Horn–Schunck** : Cette méthode permet l'optimisation d'une fonction basée sur les résidus de la contrainte de luminosité, et d'un terme de régularisation particulière exprimant le lissement attendu du champ de flux, cette méthode est détaillée par la suite [95].
* **Méthode Buxton** – Cette méthode est basée sur un modèle du mouvement des contours dans les séquences d'images [99].
* **Méthode Black-Jepson** – Calcule le flux optique par la méthode de corrélation [13].

4.4.3 Estimation de flux optique par la méthode Horn–Schunck

La méthode *Horn-Schunck* [95] d'estimation du flux optique est une méthode globale qui introduit une contrainte globale de lissage pour résoudre le problème d'ouverture. L'algorithme proposé par Horn-Schunck suppose un lissage du flux sur la totalité de l'image. Ainsi, il essaie de minimiser les distorsions de flux et préfère les solutions plus lisses.

Le flux est formulé comme une fonction énergétique globale que l'on cherche ensuite à minimiser. Cette fonction est donnée pour les flux d'images bidimensionnelles comme :

$$E = \iint \left[(I_x V_x + I_y V_y + I_t)^2 + \alpha^2(\|\nabla V_x\|^2 + \|\nabla V_y\|^2)\right] \mathrm{d}x\mathrm{d}y \qquad (4.7)$$

Où :

I_x, I_y etI_t sont les dérivées des valeurs d'intensité de l'image le long des dimensions x, y et le temps respectivement.

Le vecteur $\vec{V} = (V_x, V_y)^\top$ représente le flux optique.

Le paramètre α est une constante de régularisation. Des valeurs plus élevées de α conduisent à un flux plus lissé.

Cette fonction peut être minimisée en résolvant les équations multidimensionnelles d'Euler-Lagrange associées. Il s'agit de :

$$\begin{aligned} I_x(I_xV_x + I_yV_y + I_t) - \alpha^2\Delta V_x = 0 \\ I_y(I_xV_x + I_yV_y + I_t) - \alpha^2\Delta V_y = 0 \end{aligned} \tag{4.8}$$

Où les indices désignent à nouveau la différenciation partielle et $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ désigne l'opérateur de Laplace. En pratique, le Laplacien est approximé numériquement en utilisant des différences finies.

$$\begin{aligned} (I_x^2 + \alpha^2)V_x + I_xI_yV_y = \alpha^2\overline{V_x} - I_xI_t \\ I_xI_yV_x + (I_y^2 + \alpha^2)V_y = \alpha^2\overline{V_y} - I_yI_t \end{aligned} \tag{4.9}$$

Qui est linéaire en V_x et V_y et peut-être résolu pour chaque pixel de l'image. Cependant, comme la solution dépend des valeurs voisines du champ d'écoulement, elle doit être répétée une fois que les voisins soient mis à jour. Le schéma itératif suivant est obtenu :

$$V_x^{k+1} = \overline{V_x}^k - \frac{I_x(I_x\overline{V_x}^k + I_y\overline{V_y}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2}$$
$$V_y^{k+1} = \overline{V_y}^k - \frac{I_y(I_x\overline{V_x}^k + I_y\overline{V_y}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \quad (4.10)$$

Où k représente le numéro d'itération V_x^0 et V_y^0 représentent les estimations initiales de la vitesse qui sont initialisées à zéro.

4.4.4 Applications de flux optique

La recherche dans les littératures de la vision par ordinateur montre de nombreuses applications pouvant exploiter la notion de flux optique. nous citons cidessous des exemples de ces applications :

— **La compression des vidéos** : Cette méthode consiste à réduire la quantité de données pour réduire les coûts de stockage et de transmission des fichiers vidéo. Exemples de formats de codage vidéo : MPEG-2 Partie 2, MPEG-4 Partie 2, MPEG-4 Partie 2, H.264 (MPEG-4 Partie 10), HEVC, Theora, RealVideo RV40, VP9 et AV1. Le livre [74] traite ces algorithmes d'estimation de mouvement, en parlant de leur complexité, de leur implémentassions, de leurs avantages et de leurs inconvénients dans le domaine de la compression d'image.

— **Domaine de la robotique** : Le « flux optique » est également utilisé dans le domaine de la robotique, englobant les techniques connexes du traitement d'images et du contrôle de la navigation, y compris la détection de mouvements, la segmentation d'objet, l'information sur le temps de contact, l'accent des calculs d'expansion, la luminance et l'encodage compensé par les mouvements [3, 13, 37].

— **La vision industrielle** : L'application du flux optique permet aussi de déduire non seulement le mouvement de l'observateur et des objets dans la scène, mais aussi la structure des objets et de l'environnement. Puisque

la connaissance du mouvement et la production de schémas mentaux de la structure de notre environnement sont des composantes essentielles de la vision animale (et humaine), la conversion de cette capacité naturelle en capacité pour les ordinateurs est aussi cruciale dans le domaine de la vision industrielle [29].

4.5 Extraction des caractéristiques des mouvements

Le processus proposé pour l'extraction des caractéristiques des mouvements de l'objet humain est illustré par le diagramme 4.5. Nous obtiendrons des informations sur la vitesse (*Amplitude*) et la direction (*Angle*) du mouvement par pixel après cette étape.

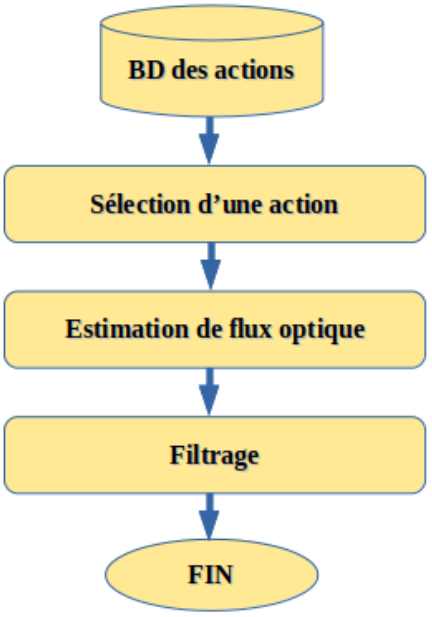


Figure 4.5 – Un processus d'extraction des caractéristiques des mouvements par l'estimation de flux optiques.

Ces étapes peuvent être expliquées comme suit :

— **Sélection d'une action** :Cette étape permettant la sélection des vidéos représentant des actions simples une par une à partir de la base de données déjà créée pendant l'étape de prétraitement.

— **Estimation de flux optique** :pour chaque triplé (les 3 trames consécutives désignées respectivement par *previous*, *current* et *next*) de l'action sélectionnée, le flux optique est calculé pour chaque pixel de la trame via l'estimation de flux optique comme suit :

$$v_{prev} = Optical\_Flow(previous, current) \tag{4.11}$$

et

$$v_{next} = Optical\_Flow(current, next) \tag{4.12}$$

tel que :

* v_{prev} et v_{next} les résultats de calcul de flux optique entre les trame (*previous*, *current*) et (*current*, *next*) respectivement.
* $Optical\_Flow$ est une fonction de calcul de flux optique. Nous avons utilisé la méthode de ***Horn–Schunck*** [95] pour l'estimation de flux optique.

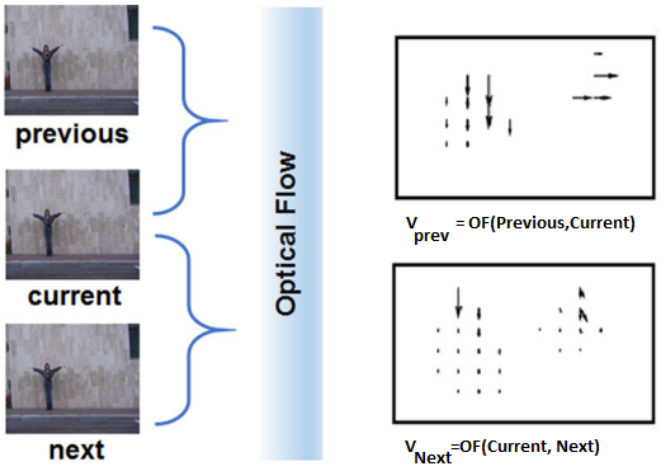


Figure 4.6 – Estimation de flux optique pour un triplé (OF : la fonction de l'estimation de flux optique entre deux trames successives).

— **Débruitage (Filtrage)** : Afin de ne conserver que les caractéristiques informatives, le processus de seuillage est utilisé, pour cela on utilise l'amplitude

du flux optique ou la vitesse de mouvement pour filtrer les valeurs inférieures à la valeur de $\tau = 0.5$, Cette valeur a été obtenue par des expériences empiriques.

4.6 Descripteur de mouvement proposé

En vision par ordinateur, les descripteurs visuels sont la description des caractéristiques visuelles du contenu des images ou des vidéos. Ils décrivent des caractéristiques élémentaires ou la combinaison de ses caractéristiques telles que la forme, la couleur, la texture ou le mouvement, et autres selon un objectif bien visé. Cette étape de construction du descripteur est la succession logique de l'étape de l'extraction des caractéristiques précédentes. En même temps, cette étape est une étape critique et la prochaine étape qui est l'étape de la classification dépend entièrement du succès de cette phase. La méthode proposée consiste à représenter une action humaine en fonction d'un ensemble de caractéristiques extraites à partir des flux optiques calculés entre trois trames successives. On récupère à partir du flux optique la vitesse de mouvement des pixels. Cela nous permet de connaître la direction (*Angle*) ainsi que sa vitesse du mouvement (*Amplitude*). Grâce à ces deux informations et la séquence des trames, nous avons pu créer un descripteur pour le mouvement. Le processus de la création d'un descripteur de mouvement fondé sur le flux optique n'est pas nouvel et il existe plusieurs descripteurs qui s'articulent sur les informations extraites à partir des flux optiques. La différence entre eux est la manière d'utilisation des informations extraites à partir du flux optique pour leur création. Dans ce que suit, nous allons parler de certains de ces descripteurs. Puis, nous parlerons de notre descripteur proposé

4.6.1 Descripteur à base de flux optique

L'utilisation du flux optique a été considérée comme une caractéristique de bas niveau dans diverses applications basées sur la vision. En effet, les caractéristiques basées sur le mouvement sont considérées comme une forte indication visuelle de la perception des scènes [102].

— **Le descripteur Histogram of Oriented Optical Flow (HOOF)** : ce descripteur est proposé par Chaudhry *et al.* [39] indiquant son invariance par rapport aux orientations de mouvement et à l'échelle. Le descripteur est construit en estimant les caractéristiques de flux optique sur chaque image sans recourir à une segmentation de l'arrière-plan ou à la localisation du sujet. Par la suite, les noyaux de Binet-Cauchy sont utilisés pour faire correspondre les histogrammes non linéaires. Son approche a été évaluée sur le jeu de données Weizmann avec un taux de classification de 95,66%.

— Martinez *et al.* [153] ont déployé le flux optique afin d'estimer la vitesse de chaque pixel. Pour chaque image, un histogramme local accumulé est construit, contenant les orientations de mouvement des vecteurs optiques discrétisés uniformément dans 32 directions. L'histogramme global de l'action humaine est composé de 192 classes concaténant 6 histogrammes consécutifs locaux. Sur la base de données Weizmann, le score de classification correcte a atteint 95% en utilisant le classifieur SVM pour la classification.

— Wang *et al.* [214] ont présenté l'image du flux optique comme une représentation ordonnée et compacte des données du flux optique à partir de trames consécutives. Sur la base de données UCF101, le score de classification correcte a atteint 89.41% en utilisant le classifieur Deep learning du type CNN pour la classification.

— Colque *et al.* [44] ont proposé les histogrammes du descripteur d'orientation et d'amplitude du flux optique par l'estimation du flux optique à partir de régions cuboïdes en tenant compte des dimensions temporelles et spatiales. Leur descripteur proposé a été appliqué à la détection d'activités anormales dans des scénarios de surveillance. L'expérimentale a été réalisé sur la base de données UCSD des activités anormales, le score de classification a atteint 71.50% en utilisant le classifieur SVM.

4.6.2 Construction du descripteur proposé

Dans cette section, nous discuterons en détail la manière que nous avons adoptée pour la construction de notre descripteur de la représentation des actions humaines.Le diagramme 4.7 représente l'algorithme de construction de ce

descripteur. Cet algorithme est basé sur l'estimation de flux optique calculé pendant l'étape de l'extraction des caractéristiques. Ce processus comprend plusieurs étapes :

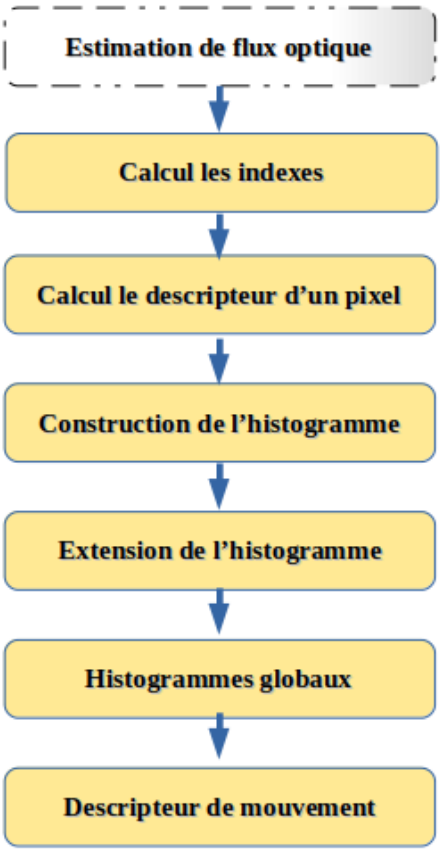


Figure 4.7 – Construction le descripteur de mouvement proposé.

4.6.2.1 Calcul des indexes

Afin de produire un descripteur basé sur un histogramme, l'orientation des vecteurs de flux optique est discrétisée en divisant le système des coordonnées polaires en 8 sections numérotées de 1 à 8. À la base de la valeur de l'angle trouvée de vecteur du flux optique, l'indexe *index* généré représente le numéro de l'un des huit secteurs où l'angle calculé appartient à cette section. Formellement cet indexe est exprimé par l'équation (4.13).

$$index_{a,b}(x,y) = \lfloor \frac{Angle_{a,b}(x,y) \times 8}{2 \times \pi} \rfloor + 1 \tag{4.13}$$

Tel que $\lfloor \ \rfloor$ représente la partie entière d'un nombre réel. a et b sont deux trames successives.

Remarque : Dans les cas où il n'y a pas de mouvement ou si le vecteur est filtré pendant le processus de seuillage, la valeur de l'*index* est mise à zéro.

4.6.2.2 Descripteur d'un pixel

Lorsque deux matrices de flux optique sont calculées pour un triplé de trames, les *index* résultant de chaque point de la trame précédente et de la trame suivante sont assemblées pour produire un nombre en base 9 qui est ensuite converti en base 10, comme exprimés en (4.14). Le nombre obtenu est la valeur de descripteur pour un point de coordonnés (x, y). La figure 4.8 illustre la manière de calcul de descripteur d'un point.

$$desc_t(x,y) = index_{t_1,t_2}(x,y) + index_{t_2,t_3}(x,y) * 9 \tag{4.14}$$

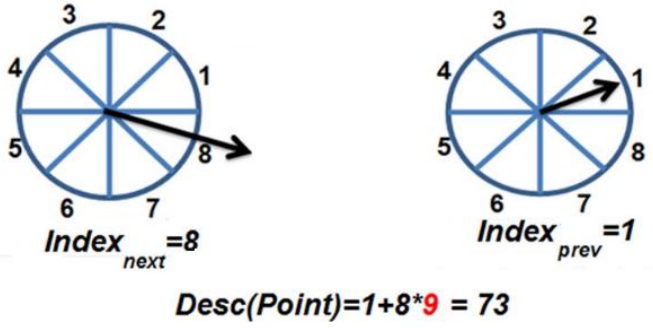


Figure 4.8 – le descripteur d'un pixel.

Sachant que : t_1, t_2 et t_3 la première, la deuxième et la troisième trame d'un triplé t. Le nombre produit à l'aide de la fonction $desc_t(x, y)$ est la valeur du descripteur pour le point de coordonnés (x, y).

4.6.2.3 Construction d'un histogramme

Sur la base d'expériences empiriques, une action simple ou atomique peut être suffisamment représentée à l'aide d'un ensemble de 15 trames consécutives dans le cas où les vidéos sont enregistrées avec une fréquence d'images de 25 trames/seconde. Le processus de codage consiste à inclure sept triplés pour chaque action humaine afin de produire l'histogramme d'orientation du flux optique exprimé dans (4.15).

$$H_t(i) = \sum_{x,y} \begin{cases} 1 & \text{if } desc_t(x,y) == i \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

Il y a une fonction de base booléenne qui renvoie 1 pour les vrais cas et 0 sinon. H_t fait référence à l'histogramme obtenu au t^{me} triplé.

Le diagramme 4.9 montre le processus de construction d'un histogramme d'un triplé.

La figure 4.10 illustre la procédure complète pour générer l'histogramme des propriétés d'orientation du flux optique à partir des caractéristiques locales.

Dans la pratique, on a calculé 80 caractéristiques pour chaque histogramme d'un triplé. Nous proposons pour coder une action de 15 trames un descripteur local de 640 caractéristiques différentes, selon l'équation 4.16 suivante :

$$D = [H_1 \ ... \ H_7 \ \sum_{t=1}^{7} H_t] \tag{4.16}$$

4.6.2.4 Extension de l'histogramme

Pendant cette étape, un certain nombre de caractéristiques de mouvement qui peuvent intégrer des traits plus distinctifs à l'action humaine sont construites en appliquant des techniques de fusion simples telles que des méthodes arithmétiques et statistiques exécutées sur l'ensemble des histogrammes d'orientation produits à partir de (4.15). Dans (4.17), l'équation exprime le vecteur de caractéristique

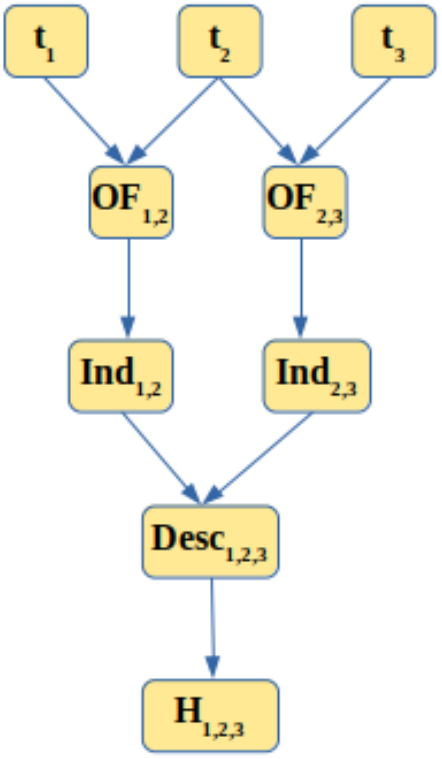


FIGURE 4.9 – Construction de l'histogramme d'un triplé.(**t** :trame, **OF** : fonction de calcul de flux optique, **Ind** : fonction de calcul les indexes entre deux trames, **Desc** : Calcul le descripteur d'un triplé et **H** : histogramme de descripteur d'un triplé)

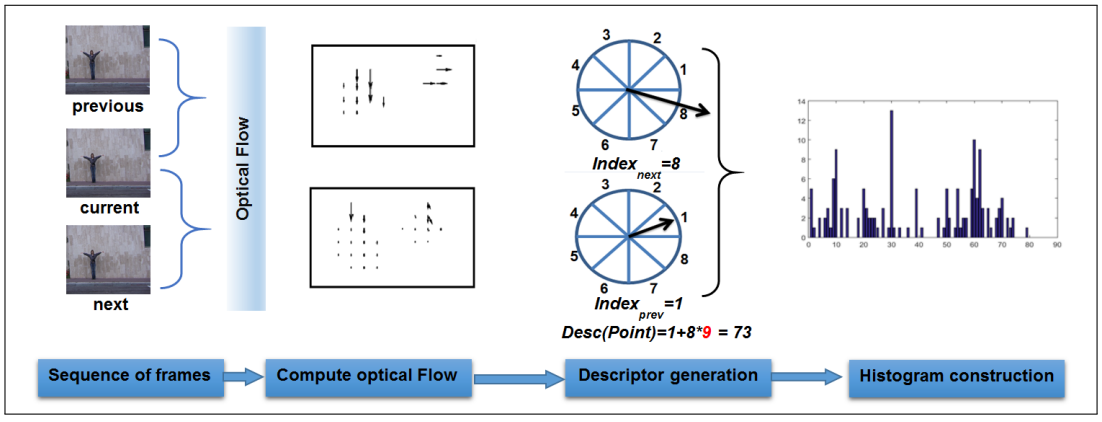


FIGURE 4.10 – Construction d'histogrammes à partir de descripteurs de mouvement locaux

produit en concaténant les différents histogrammes où l'écart-type est abrégé en STD. Le vecteur d'action résultante est composé de caractéristiques décrivant uniquement des caractéristiques locales basées sur la dynamique sans tenir compte des informations relatives à la structure spatiale globale de l'activité ni des données anthropométriques et anatomiques.

$$F = [H_1 \ ... \ H_7 \ \ Mean(H_1 \ ... \ H_7) \ \ STD(H_1 \ ... \ H_7) \ \ \sum_{t=1}^{7} H_t] \tag{4.17}$$

Dans la pratique, nous avons compté après cette étape d'extension 951 caractéristiques différentes.

4.6.2.5 Caractéristiques globales

Pour extraire les caractéristiques spatiales globales qui décrivent mieux les propriétés géométriques des signaux de mouvement, chaque image du flux optique prise à partir de deux images consécutives est extraite à la fois verticalement et horizontalement dans des barres adjacentes de largeur similaire, comme le montre la figure 4.11. Contrairement à la plupart des études qui consistent à scinder la région d'intérêt en une grille de cellules liée à la localisation du sujet. Étant donné que les personnes peuvent se déplacer et qu'il est essentiel de capturer le déplacement spatial entre les images, deux histogrammes sont construits à partir de la séquence d'images de flux optique devant exprimer le mouvement spatial verticalement et horizontalement. Les vecteurs de flux optique contenus dans chaque barre sont accumulés ensemble dans leurs casiers respectifs d'histogramme, quelles que soient les orientations de mouvement les mieux prises en compte lors de l'extraction de caractéristiques locales.

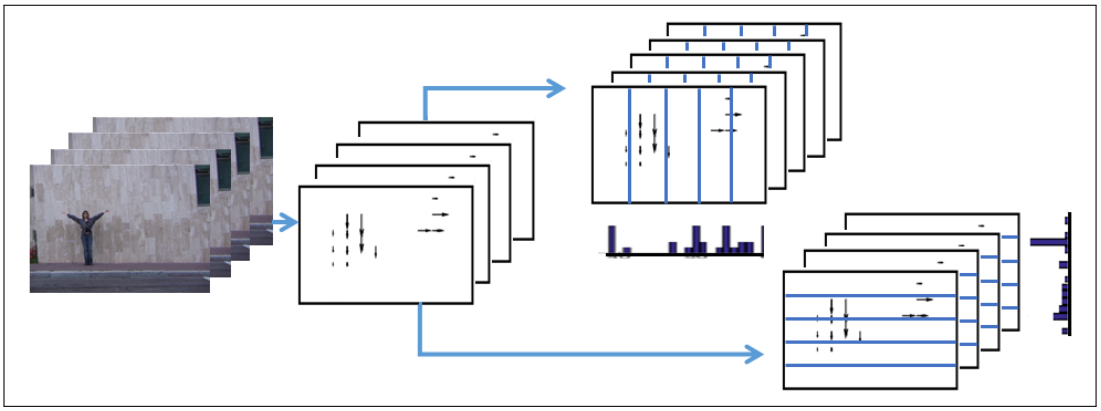


Figure 4.11 – Dérivation des caractéristiques du mouvement global

Les caractéristiques de notre descripteur peuvent être représentées par la concaténation des caractéristiques locales et globales selon l'équation 4.18 :

$$F = [F_{Local} \quad F_{Global}] \tag{4.18}$$

Nous avons compté 1015 et 1035 caractéristiques selon le nombre de divisions 10 et 20 respectivement.

4.7 Conclusion

Dans ce chapitre, nous avons discuté la manière de la construction de notre descripteur de mouvement proposé. Dans le même contexte, nous avons également parlé du flux optique, les méthodes d'estimation de mouvement et la méthode de *Horn-Chuck* que nous avons adopté pour l'extraction des caractéristiques cinématiques locales et globales. Celle-ci est l'élément de base pour la construction de notre histogramme et par la suite notre descripteur. Cette méthode de construction permet la segmentation de l'objet en mouvement sans passer par l'étape de soustraction de l'arrière-plan. La question que nous devons poser est la suivante : "comment utiliser efficacement ce descripteur pour classifier les actions de l'être humain" et nous y trouverons une réponse dans le chapitre suivant.

5 Processus de classification proposée

5.1 Introduction

L'analyse des activités humaines à partir de séquences vidéo nécessite différents niveaux de traitement. Le traitement de bas niveau qui consiste à déceler les régions saillantes des mouvements. Le traitement de niveau intermédiaire qui comprend l'extraction de l'information visuelle et sa représentation sous une forme la plus concise possible, c'est-à-dire la construction d'un descripteur, qui est aussi invariable que possible. Enfin, un traitement de haut niveau permettant l'interprétation de ces informations et alors la reconnaissance de l'activité humaine. Il existe une multitude de techniques dans la littérature pour réaliser chacune de ces trois étapes. La reconnaissance de l'action humaine est un grand problème de classification car le dernier objectif auquel nous aspirons est d'étiqueter le mouvement par l'une des actions reconnues.

Ces dernières années, les chercheurs ont de plus en plus recours aux techniques d'apprentissage automatique, notamment dans la détection, la reconnaissance et la prédiction des mouvements humains. De ce fait, nous présentons tout d'abord ces techniques d'apprentissage afin de permettre une meilleure compréhension des autres sections de cette étude et éviter toute redondance de définitions.

Puis, nous poursuivons avec une description succincte des techniques de réduction de dimensions ou la sélection des caractéristiques. Enfin, nous allons révéler les techniques de classification classiques et modernes (Deep learning) utilisées dans nos recherches pour classifier les actions humaines et nous présentons aussi les résultats expérimentaux obtenus.

5.2 Apprentissage automatique

La classification des actions est la dernière étape de notre système proposé pour la reconnaissance des actions humaines, la réalisation de celle-ci se fait à l'aide des algorithmes d'apprentissage automatique. Nous allons présenter brièvement le concept d'apprentissage automatique et ses algorithmes bien connus. Nous allons nous concentrer sur les algorithmes que nous avons utilisés dans nos expériences de cette thèse.

Le concept de l'***apprentissage machine*** ou l'***apprentissage automatique*** (***Machine learning*** *en anglais*) est un champ de recherche de l'IA dont l'intérêt majeur est d'extraire des connaissances à partir des données à l'aide des algorithmes permettant la résolution de tâches complexes qui font partie des tâches exclusives des humains (Apprentissage statistique). À la différence des algorithmes classiques, les algorithmes d'apprentissage automatique permettent aux systèmes à la fois d'*apprendre* et de *raisonner*, c'est la notion de l'intelligence (Figure 5.1).

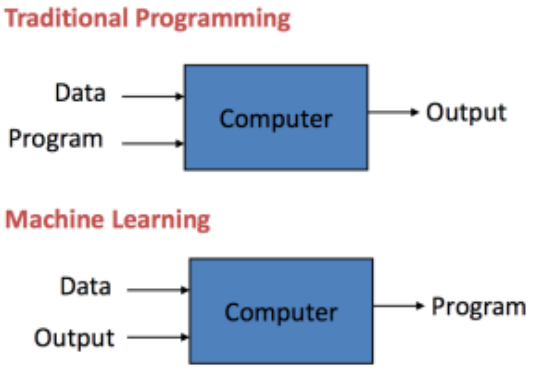


Figure 5.1 – Programmation traditionnelle VS Apprentissage automatique. Source de l'image :https://machinelearningmastery.com/basic-concepts-in-machine-learning/

Pour être intelligent, un système évoluant dans un environnement changeant devrait avoir la capacité d'améliorer ses performances à partir de données acquises en cours de fonctionnement et ainsi apprendre de son expérience. Partant de ce fait, les techniques d'apprentissage machine ont été développées de sorte à modéliser l'apprentissage de point de vue mathématique afin de générer un modèle optimisant un critère de performance et d'analyser de manière automatique un ensemble limité de données représentant une tâche précise : ***phase d'entraînement***, en vue d'être appliqués sur de nouvelles données : ***phase de test***. Le modèle peut être prédictif pour prévoir des valeurs futures, descriptif pour acquérir des connaissances et détecter des schémas à partir des données, ou les deux (Figure 5.2).

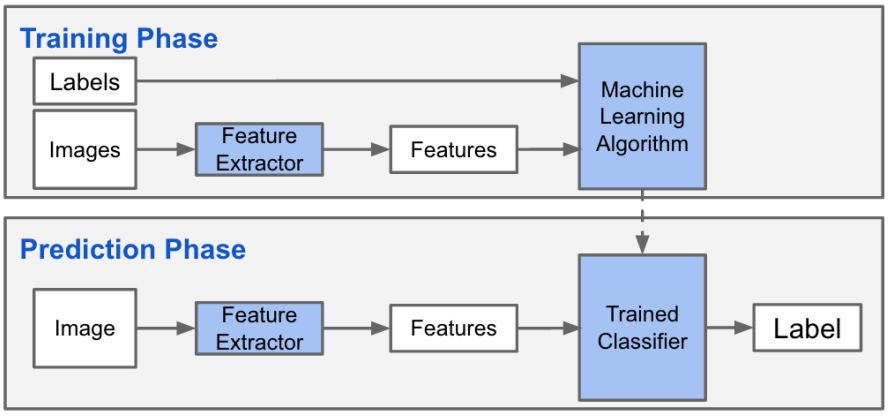


Figure 5.2 – Phases d'apprentissage automatique. source de l'image :http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/

Dans la suite, nous présentons les quatre classifieurs que nous avons utilisés dans la partie expérimentale de notre thèse. Il s'agit des classifieurs KNN, arbre de décision et SVM et aussi le Deep learning. Il ne s'agit pas de faire une présentation exhaustive de ces algorithmes mais seulement de préciser leurs fonctionnements. Par souci de concision, chaque algorithme sera décrit brièvement, nous concentrons seulement sur les points importants pour faciliter la compréhension de leurs principes.

5.2.1 k-plus proches voisins

Nous pouvons dire que le principe de cet algorithme peut être exprimé comme suit :"***Tell me who your neighbors are, and I'll tell you who you are***". L'algorithme de k-plus proches voisins *k-ppv* (k-nearest neighbor en Anglais (*k-NN*) où simplement *KNN*)[8] se base sur une comparaison directe entre le vecteur caractéristique représentant l'entité à classer et les vecteurs caractéristiques représentant des entités de référence (Figure 5.3)

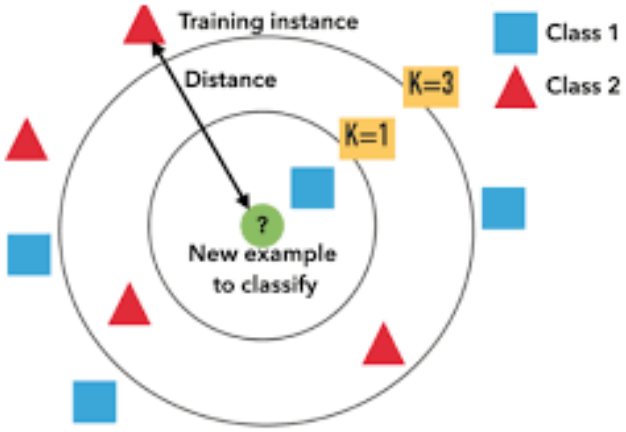


Figure 5.3 – Le principe de KNN. (Source :https://informatic-ar.com/k-nearest-neighbor-algorithm/)

L'algorithme de *KNN* est utilisé pour la classification et la régression. Dans les deux cas, l'entrée est constituée d'un ensemble d'apprentissage et l'entier k pour les voisins les plus proches. La sortie dépend si cet algorithme est utilisé pour la classification ou la régression :

— Dans le cas de la classification, la sortie est l'une des classes du problème traité. Un objet est classé par un vote majoritaire de ses voisins, l'objet est affecté à la classe la plus commune parmi ses k voisins les plus proches. Si $k = 1$, alors l'objet est simplement affecté à la classe de ce voisin le plus proche.

— En régression, la sortie est la propriété de l'objet. Cette valeur est la moyenne des valeurs de k voisins les plus proches.

5.2.1.1 Principe de fonctionnement

L'algorithme ci-dessous montre le mécanisme de fonctionnement de KNN pour prédire la classe d'une nouvelle entrée :

Entrées :

— Un ensemble de données D .

— Une fonction de définition de distance d.

— Un nombre entier $K \geq 1$

Sortie : La prédiction y de la nouvelle observation X.

1. Calcule toutes les distances de l'observation X avec les autres observations du jeu de données D à l'aide de la fonction de définition de distance d
2. Retenir les K observations du jeu de données D les proches de X
3. Prendre les valeurs de y des K observations retenues :
   - ∗ Dans le cas de la régression, calculer la moyenne (ou la médiane) de y retenues
   - ∗ Si on effectue une classification , y est la classe majoritaire de ses voisins
4. Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par KNN pour l'observation X.

5.2.2 Arbre de décision

Un ***arbre de décision*** ou un ***arbre de classification*** est un algorithme d'apprentissage automatique qui génère des arbres de décision à partir des données d'entraînement pour résoudre les problèmes de classification ou de régression. L'idée de cet algorithme est de réaliser la classification ou la régression d'un objet par une suite de tests sur les attributs (caractéristiques) qui le décrit. Ces tests sont organisés de telle façon que la réponse à l'un d'eux indique à quel prochain test doit-on soumettre cet objet. On peut résumer la définition de l'arbre de décision par :

Un arbres de décision est un classifieur pour des entités représentées dans un formalisme attribut/valeur tel que :

— Les nœuds de l'arbre testent les attributs

— La branche représente la valeur de l'attribut testé

— Les feuilles indiquent les catégories (classes)

Exemple :

La figure 5.4 montre un exemple d'arbre de décision. Les attributs sont "Fièvre", "Douleur" et "Toux". Les classes sont les maladies : "Appendicite","Rhume","Mal de gorge","Refroidissement" et "Rien".

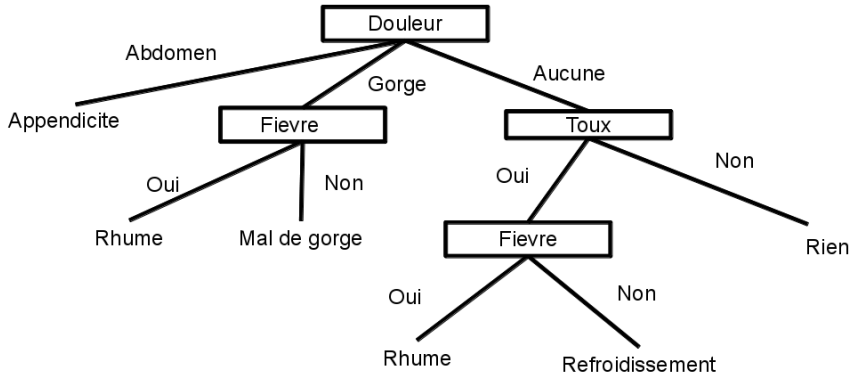


Figure 5.4 – Un exemple d'arbre de Décision.(source :http://www.up2.fr/M1/td/TD10_2.html)

5.2.3 Machine à Vecteur de Support (SVM)

Les ***machines à vecteurs de support*** ou ***séparateurs à vaste marge*** (Support Vector Machine : SVM en Anglais) sont un ensemble de méthodes d'apprentissage supervisé utilisées pour la classification et la régression. L'algorithme SVM original a été inventé par Vladimir Vapnik et la version standard actuelle nommée "soft margin" a été proposée par *Corinna Cortes et Vladimir Vapnik* [46]. Les SVM sont une généralisation des classifieurs linéaires. Elles s'appliquent aussi

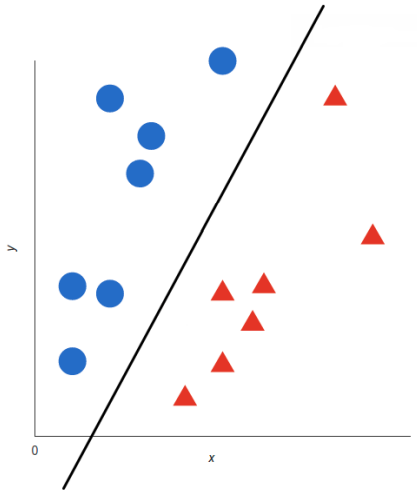


Figure 5.5 – Problème de séparation à deux classes

bien à des ***problèmes linéairement séparables*** ou ***non séparables***. Intuitivement, un modèle SVM est une représentation des exemples sous forme des points dans l'espace schématisé de sorte que les exemples des différentes classes soient séparés par une espace (hyperplane) aussi large que possible. Dûs de leurs avantages, les SVM sont très utilisées pour résoudre des nombreux problèmes pratiques de classification tels que : La classification des données biologiques et physiques, la reconnaissance des expressions faciale, la classification de textures, le E-learning, la reconnaissance des images et des vidéos basées sur le contenu, etc.

5.2.3.1 Principe

Supposant un ensemble d'entraînement : $\{(x_i, y_i)\}_{i=1,\ldots,n}$ où $x_i \in \mathcal{X}$ et $y_i \in \{-1, +1\}$ (cas de deux classes). L'objectif est de construire une fonction $f : \mathcal{X} \to \mathbb{R}$ qui permet de prédire si un nouvel $x \in \mathcal{X}$ appartient à la classe -1 ou à la classe $+1$.

On cherche alors une "surface de séparation" (Figure 5.5) tel que si $f(x) > 0$ alors x est affecté à la classe $+1$ et sinon x est affecté à la classe -1.

Séparateurs à vaste marge : Pour un problème de classification linéaire on suppose que les deux classes (-1 et $+1$) sont séparables par un hyperplan, la fonction f a donc la forme :

$$f(x) = \sum_{i=1}^{n} w_i x_i + b = \langle w, x \rangle + b \tag{5.1}$$

Où w est le vecteur orthogonal à l'hyperplan et b est le déplacement par rapport à l'origine.

La "marge" est définie comme la distance entre le plus proche exemple d'apprentissage et l'hyperplan de séparation. Pour un hyperplan H on a :

$$\text{Marge}(H) = \min_{x_i} d(x_i, H) \tag{5.2}$$

Les SVM linéaires cherchent le séparateur qui maximise la marge. C'est la raison de la nomination "séparateur à vaste marge".

Dans le cas : SVM linéaire (cas séparable)

$$\text{Marge} = \frac{2}{||w||} \tag{5.3}$$

On arrive à un problème d'optimisation suivante (Figure 5.6) :

$$\begin{cases} \min_{w,b} \frac{1}{2}||w||^2 \\ \text{tel que } y_i(w \cdot x_i + b) \geq 1, i = 1, \ldots, n \end{cases} \tag{5.4}$$

Remarque : Il existe d'autres types de SVM tel que : SVM avec les données non séparables linéairement, SVM multi-classe et SVM pour la régression mais ce n'est pas l'essentiel de cette thèse.

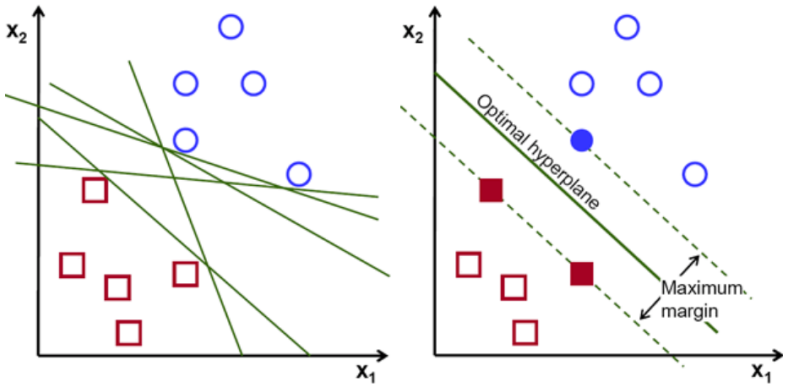


Figure 5.6 – Principe du Séparateur à Vaste Marge (SVM). source de l'image : https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f

5.2.4 Deep learning

L'*apprentissage profond*, *apprentissage approfondi* ou *deep learning* et aussi deep *structured learning* ou *hierarchical learning* est une classe d'algorithmes d'apprentissage automatique[56], leurs fonctionnements reposent sur les points suivants :

— Utilise une cascade de couches multiples de traitements non linéaires pour l'extraction et la transformation des caractéristiques. Chaque couche suivante utilise la sortie de la couche précédente comme entrée.

— Apprendre de façon supervisée (p. ex., classification) et/ou non supervisée (p. ex., analyse de modèles).

— L'apprentissage à travers les différentes couches ; les caractéristiques passent de bas niveau à de plus haut niveau, où les différents niveaux correspondent à différents niveaux d'abstraction des données (Figure 5.7).

Des architectures de deep learning telles que les *deep neural networks*, *deep belief networks* et *recurrent neural networks* ont été appliquées à des domaines tels que la vision par ordinateur, la reconnaissance vocale, le traitement du langage naturel, la reconnaissance audio, le filtrage des réseaux sociaux, la traduction automatique, la bio-informatique, la conception de médicaments, l'analyse des images médicales, l'inspection des matériaux et les jeux de société, qui ont donné des résultats comparables et parfois supérieurs aux experts humains [43, 126, 187].Pendent

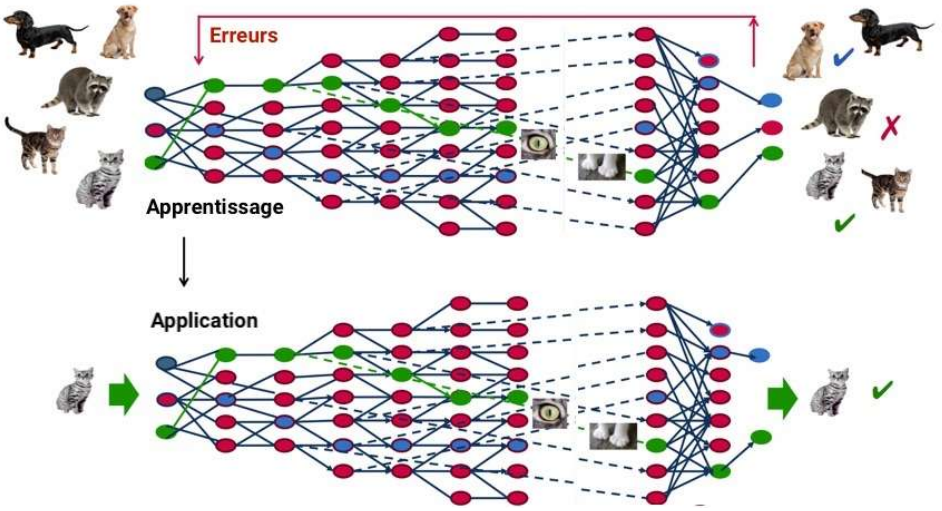


FIGURE 5.7 – Le principe de Deep learning : à chaque couche du réseau neuronal correspond un aspect particulier de l'image. (Source : https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/)

les années 2000, ces progrès ont stimulé des investissements privés, universitaires et publics importants, notamment de la part des GAFA (Google, Apple, Facebook, Amazon).

Les modèles d'apprentissage en profondeur s'inspirent vaguement des modèles de traitement de l'information et de communication des systèmes nerveux biologiques, mais ils présentent diverses différences par rapport aux propriétés structurelles et fonctionnelles des cerveaux biologiques (en particulier du cerveau humain), ce qui les rend incompatibles avec les preuves en neurosciences [152, 167, 14].

5.2.4.1 Autoencoder

Nous avons utilisé dans notre phase expérimentale un classifieur deep learning du type ***Autoencoder*** ou ***Auto-encodeur***. C'est un algorithme non supervisé qui permet de construire une nouvelle représentation de données en entrée (Figure 5.8).

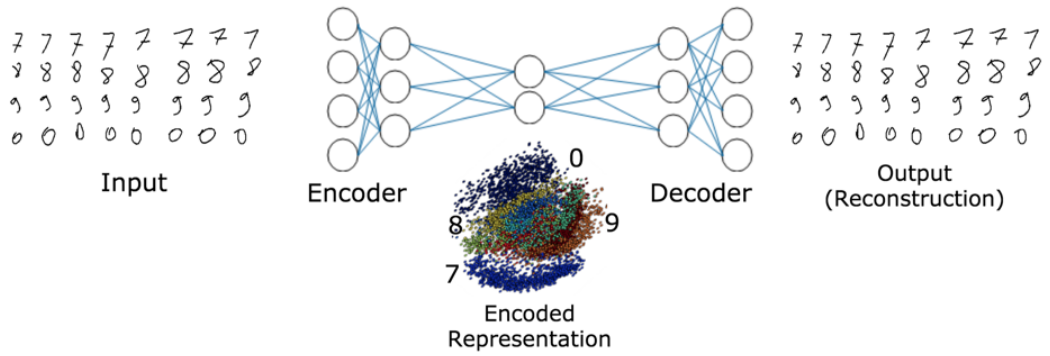


FIGURE 5.8 – Autoencoder. (Source :https://dataanalyticspost.com/Lexique/auto-encodeur/)

Il se constitue de deux parties : l'***encodeur*** et le ***décodeur***. L'encodeur est constitué d'un ensemble de couches cachées de neurones, qui traitent les données afin de construire des nouvelles représentations. En revanche, le décodeur reçoit ces représentations et les traite afin d'essayer de reconstruire les données de départ. La mesure d'erreur commise par l'auto-encodeur se fait par l'étude des différences entre les données reconstruites et les données initiales. L'entraînement consiste à modifier les paramètres de l'auto-encodeur afin de réduire l'erreur de reconstruction mesurée sur les différents exemples de la base de données.

5.3 Sélection des caractéristiques (Features selection)

De nombreux facteurs influent sur le succès de l'apprentissage automatique dans une tâche précise, en particulier la représentation et la qualité des données. Théoriquement, avoir plus de caractéristiques devrait conduire à une discrimination plus grande, mais l'expérience pratique avec les algorithmes d'apprentissage automatique a montré que ce n'est pas toujours le cas et parfois, cela affecte négativement la qualité des résultats [90]. Il pourrait que l'ensemble de caractéristiques ne sont pas toutes pertinentes. Il est possible que certaines soient peu informatives, redondantes ou même inutiles au système pour l'accomplissement de sa tâche. D'après *Ditterrich* [61], les performances du système de classification dépendent fortement des relations entre le nombre d'échantillons utilisés, le nombre

de propriétés examinées et la complexité du système. Dans notre cas, Nous proposons deux modes pour le processus de classification ; avec /sans sélection des caractéristiques. En conséquence, nous avons donnée un aperçu sur ce concept, et sur l'algorithme que nous avons choisi pour la sélection des caractéristiques à partir de notre descripteur.

5.3.1 Définitions

Selon *Chouaib* [42], la sélection de caractéristiques est définie comme un processus de recherche permettant de trouver un sous-ensemble "pertinent" de caractéristiques parmi celles de l'ensemble de départs. La notion de pertinence d'un sous-ensemble des caractéristiques dépend toujours des objectifs et des critères du système. En général, le problème de sélection de caractéristiques peut être défini par :

Soit $F = \{f_1, f_2, \ldots, f_N\}$ un ensemble des caractéristiques de taille N où N représente le nombre total des caractéristiques étudiées.

Soit Ev une fonction qui permet d'évaluer un sous-ensemble des caractéristiques. Nous supposons que la plus grande valeur de Ev soit obtenue pour le meilleur sous-ensemble des caractéristiques.

L'objectif de la sélection des caractéristiques est de trouver un sous-ensemble F' $(F' \subseteq F)$ de taille $N'(N' \leqslant N)$ tel que :

$$Ev(F') = \max_{Z \subseteq F} Ev(Z) \tag{5.5}$$

Où $|Z| = N'$ et N' est, soit un nombre prédéfini par l'utilisateur, soit à contrôler par une des méthodes de génération de sous-ensembles.

5.3.2 Algorithmes de sélections de caractéristiques

Frédéric Grandier dans ça thèse [87] a détaillé les différentes techniques permettant de mettre en oeuvre les différents modules de sélection des caractéristiques et la figure 5.9 illustre ce processus de sélection des caractéristiques :

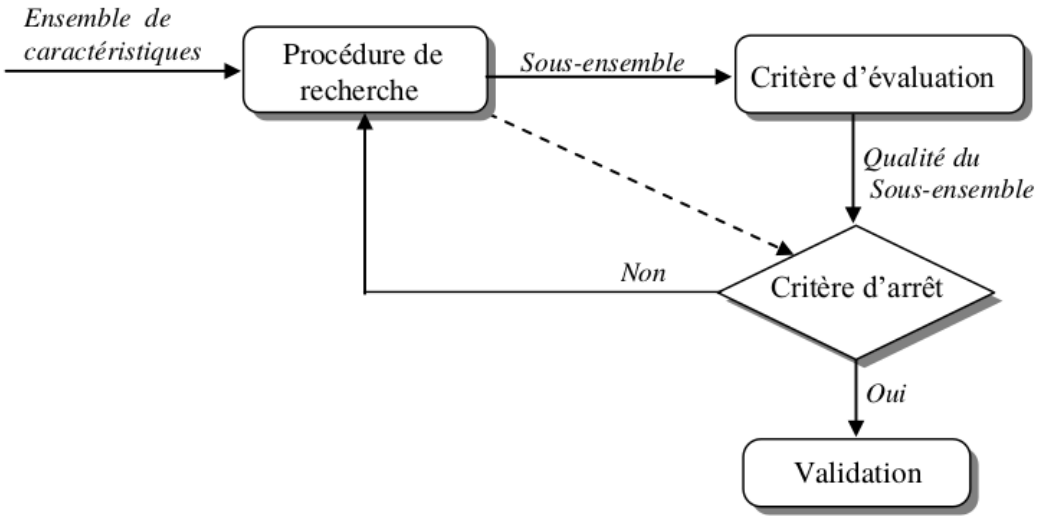


Figure 5.9 – Processus de sélection des caractéristiques [87].

— ***Procédures de recherche :*** Le rôle de cette procédure consiste à générer les sous-ensembles de caractéristiques qui seront évalués. Si l'ensemble de départ contient n caractéristiques, le nombre total de sous-ensembles possibles est $2^n - 1$. Il existe différentes techniques permettent de générer les sous-ensembles candidats.

— ***Évaluation de caractéristiques :*** Les chercheurs du domaine énoncent que le meilleur critère d'évaluation est le taux de reconnaissance de la classification elle-même, mais cette mesure est rarement fiable du fait qu'elle est obtenue à partir d'un nombre fini et limité d'échantillons (Problème de généralisation). Il est également important de mentionner qu'un sous-ensemble de caractéristiques est optimal uniquement par rapport au critère d'évaluation. De ce fait, le choix de ce critère est très important.

— ***Critère d'arrêt :*** Une fois le critère d'évaluation des caractéristiques et la méthode d'évaluation définie, tous les sous-ensembles proposés sont évalués et celui retenu est le plus pertinent dans le sens du critère d'évaluation. Dans le cas des méthodes de sélection séquentielles, le critère d'arrêt

est fortement conditionné par la mesure de pertinence des caractéristiques. La recherche est arrêtée lorsqu'aucune des caractéristiques restantes n'est considérée comme pertinente. La pertinence d'une caractéristique peut être obtenue en calculant des tests statistiques. Lorsque le calcul de ces tests n'est plus possible, en raison de leur complexité, une dernière solution est l'utilisation de l'heuristique. L'une des heuristiques couramment utilisées est le calcul d'une estimation de l'erreur de généralisation pour les différents sous-ensembles testés. Le sous-ensemble choisi à la fin de la procédure est bien sûr celui qui donne les meilleures performances. L'erreur de généralisation peut être calculée à l'aide d'un ensemble de validation, d'une validation croisée ou d'autres estimations.

— ***Procédure de validation :*** *Dash et Liu* [51] proposent d'ajouter un quatrième composant à un algorithme de sélection de caractéristiques : une procédure de validation. Deux alternatives sont proposées en fonction de la nature des données utilisées lors de l'exécution de cette procédure : artificielle ou réelle. En général, une base de données synthétique est créée pour tester un concept ou une application particulière. Par conséquent, les caractéristiques pertinentes sont connues et identifiées. La validation d'un algorithme sera alors directe puisqu'il suffit de vérifier si le sous-ensemble sélectionné contient les caractéristiques pertinentes. Dans le cas de données réelles, les caractéristiques pertinentes ne sont généralement pas connues, la procédure consiste alors à évaluer la précision de la classification obtenue avec le sous-ensemble de variables sélectionnées au moyen d'un classifieur (classificateur Bayes, ...).

5.3.3 Algorithme ASFFS

Les algorithmes les plus couramment utilisés dans le domaine de la sélection des caractéristiques sont connue sous le nom "les algorithmes séquentiels" sont les algorithmes du type "***Sequential Forward Selection (SFS)***". Le principe de ces algorithmes est le suivant :"*nous partons d'un ensemble vide de variables et ajoutons une variable à chaque étape à chaque fois que la performance du modèle s'améliore.*". Ces méthodes sont rapides, car leur complexité est polynomiale

($O(n)$) et très simple à mettre en œuvre. Cependant, ces méthodes sont sous-optimales parce que les variables ajoutées ne sont jamais remises en cause. En revanche, les versions flottantes de ces méthodes s'appellent " ***Sequential Forward Floating Search methods SFFS*[176]**" sont considérées comme les plus efficaces. La méthode SFFS consiste à appliquer, après chaque étape forward, x étapes backward, la valeur de x est déterminée dynamiquement : parmi l'ensemble des variables constituant le sous-ensemble courant, on enlève ces variables une par une tant que cela améliore les performances du modèle. Aucune paramétrisation de x n'est nécessaire. L'algorithme **ASFFS** (***Adaptive Sequential Forward Floating Search*[196]**) : qu'on a utilisé dans notre travail pour la sélection des caractéristiques est une version la plus sophistiquée de SFFS(Sequential Forward Floating Search). Cet algorithme est une génération des algorithmes précédents du type séquentiels flottants. Il permet de se rapprocher du sous-ensemble optimal en considérant l'ajout et le rejet des attributs par tuples et non par individus. La taille de ces tuples est déterminée de manière adaptative. Toutefois, la complexité de ces algorithmes a augmenté car elle dépend du nombre d'étapes forward et backward implémentées.

5.4 Méthodologie de classification proposée

À cette étape nous avons adopté deux modes de classification dans notre approche de reconnaissance des activités humaines. La première méthode concerne la classification des actions en utilisant les classifieurs classiques tels que KNN, Arbre de décision et SVM avec/sans le processus de la sélection des caractéristiques. La deuxième méthode traite l'utilisation de nouvelles techniques de classification appelées "Deep Leanring". La différence entre les deux méthodes réside dans le fait que la première méthode nécessite deux phases : une étape d'extraction et une autre pour la représentation des caractéristiques, tandis que la deuxième méthode n'en a pas besoin, elle nécessite seulement un prétraitement tel que la normalisation et / ou l'extraction de certaines caractéristiques (Figure 5.10).

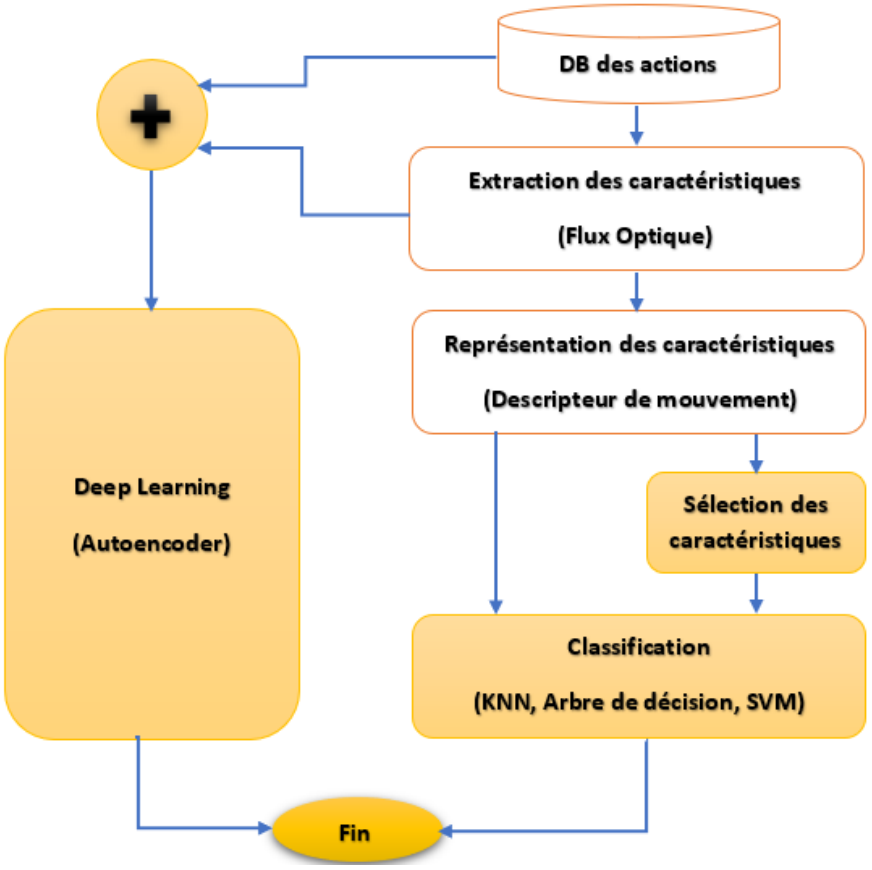


Figure 5.10 – Processus proposé pour la classification des actions

5.4.1 Classification classique

Comme nous l'avons déjà dit précédemment, ce type de classification dépend des résultats obtenus pendent l'étape de présentation des caractéristiques sous forme d'un descripteur de mouvement. Ces résultats ont été utilisés de deux manières différentes. Dans un premier temps, nous avons utilisé les caractéristiques brutes pour le processus de classification, dans le second cas, on a pris que les caractéristiques résultantes de processus de sélection des caractéristiques. Par la suite, nous avons comparé les résultats obtenus des expériences dans les deux cas. Pour cela, nous avons utilisé trois types de classifieurs classiques, KNN, arbre de décision et SVM. Notant que notre discussion sur "les classifieurs classiques" ne signifie pas que la recherche dans ce domaine a été arrêtée, il existe de nouvelles recherches visant à améliorer la qualité des résultats de la classification telle que [77, 109] .

5.4.1.1 Classification sans sélection des caractéristiques

Nous avons étudie l'étape de classification sans sélection des caractéristiques. Cela signifie que nous avons utilisé les données dérivées à partir de descripteur de manière brute sans aucun changement. Nous pouvons expliquer les raisons de ce choix par la lenteur de la méthode de sélection des caractéristiques. Nous avons appliqué les classifieurs KNN, Arbre de décision et SVM sur les mêmes jeux de données pour expliquer les points forts de l'utilisation des sélections des caractéristiques comme une étape préliminaire de l'étape de classification. Les résultats qu'on a obtenus sont bien illustré dans les paragraphes suivants.

5.4.1.2 Classification avec sélection des caractéristiques

Pour ce paradigme de classification, une procédure de sélection des caractéristiques est mise au point afin d'obtenir des caractéristiques distinctives et représentatives pour notre application de la reconnaissance des actions humaines. Comme il est impossible d'effectuer une recherche exhaustive ou par la force brute pour toutes les combinaisons de sous-ensembles afin de trouver le sous-ensemble de caractéristiques le plus discriminatoire. Ceci est dû à la dimension du vecteur de la caractéristique brute. Alternativement, l'algorithme de sélection adaptative séquentielle flottante en avant ou ***Adaptive Sequential Forward Floating Search (ASFFS)*** est utilisé pour extraire un sous-ensemble de caractéristiques (subset) (Section Algorithme ASFFS). Dans notre recherche, une fonction objective est proposée comme métrique d'évaluation qui évalue le caractère distinctif de chaque vecteur brut ou ensemble de caractéristiques afin d'extraire les caractéristiques optimales pour la reconnaissance des activités humaines. Le critère fondé sur la validation est utilisé pour sélectionner les caractéristiques représentatives qui minimisent les erreurs de classification et maximisent la séparation interclasses entre les différentes classes d'activités humaines. Une approche de filtrage avancée peut-être potentiellement employée pour optimiser la séparation entre les différentes classes [140].

Dans le cadre de notre recherche, une procédure de vote similaire à celle utilisée pour le classificateur KNN est utilisée. Le critère d'évaluation utilise un

coefficient w qui reflète l'importance des voisins les plus proches appartenant à la même classe. Une valeur de score pour une instance donnée s pour appartenir à une classe c est exprimée en (5.6). L'approche *Winner-take-all* est potentiellement utilisée dans le cadre de la procédure de sélection de la même manière pour obtenir le sous-ensemble optimal de caractéristiques ayant le score le plus élevé [207, 215].

$$P(s,c) = \frac{\sum_{i=1}^{N_c-1} z_i(s,c) w_i}{\sum_{i=1}^{N_c-1} w_i} \tag{5.6}$$

Tel que N_c est le nombre d'objets dans la classe c, et le coefficient w_i à l'i^{me} instance le plus proche est inversement lié à son voisinage :

$$w_i = (N_c - i)^2 \tag{5.7}$$

z_i est calculé comme suit :

$$z_i(s,c) = \begin{cases} 1 & \text{if } nearest(s,i) \in c \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

La fonction $nearest(s,i)$ récupère le candidat i^{me} en termes de proximité de l'instance s. Pour déduire la distance et le niveau de proximité, la distance euclidienne est calculée entre différentes instances. Pour évaluer si un sous-ensemble des caractéristiques est le plus apte à classer les activités humaines, une métrique qui est calculée à l'aide de la validation croisée "leave-one-out" est intégrée dans la procédure de sélection des caractéristiques. En termes simples, le descripteur d'activité finale avec le sous-ensemble optimal de caractéristiques est composé d'un sous-ensemble des caractéristiques parmi l'espace brut F de sorte que la valeur de validation maximale est égale à la somme moyenne de toutes les valeurs calculées pour les candidats N comme expliqué dans l'équation (5.9) :

$$Action = \arg\max_{subset \in F} \left(\frac{\sum_{x=1}^{N} L_{subset}(x)}{N} \right) \tag{5.9}$$

Tel que L est la fonction de validation croisée "leave-one-out".

5.4.2 Classification avec Deep learning

Dans le contexte de notre recherche, le deuxième paradigme de classification, l'apprentissage en profondeur est utilisé sur le même jeu de données pour évaluer les possibilités d'utilisation des caractéristiques de flux optique pour la reconnaissance des activités humaines. Dans notre article [132], nous avons proposé d'utiliser un Deep learning du type ***Autoencoder*** pour la classification des actions. Le réseau Autoencoder proposé est constitué de trois types des couches, une couche d'entrée où sa taille est égale à la taille d'un vecteur d'image des caractéristiques de mouvement en entrée, une série de trois couches cachées où le nombre de neurones décroît progressivement d'une couche à l'autre et enfin, la couche de sortie du type Softmax permet de prédire la classe de l'image de caractéristiques en entrée. Dans cette optique, nous nous basons sur l'idée d'une représentation 2D de séquences flux optique obtenu pour une action en combinant les séquences d'images en une seule image appelée Binary Motion Image (BMI). Cette image de caractéristiques est considérée comme l'entrée de notre réseau du type Autoencoder. La figure 5.11 montre l'architecture générale de ce système.

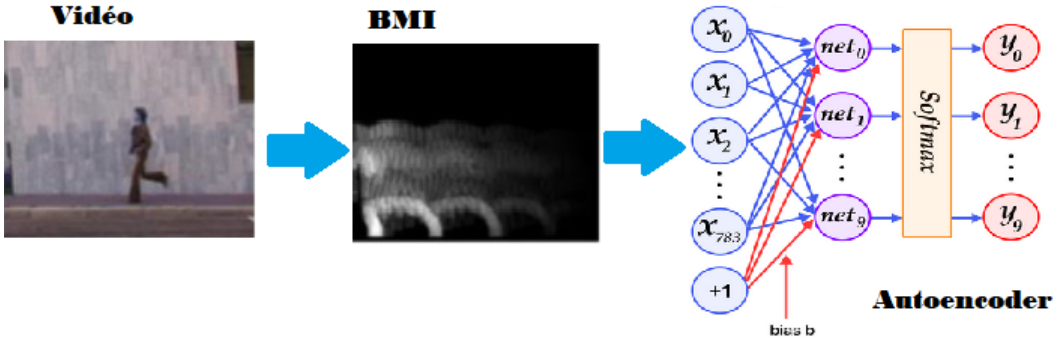


Figure 5.11 – Architecture d'un système de reconnaissance des actions huamain à base de Autoencoder(**BMI** :Binary Motion Image) [132]

La figure 5.12 indique aussi l'architecture détaillée de notre réseau de type Autoencoder.

Dans le même contexte, nous avons introduit dans l'article [82] une extension du système précédent (Figure 5.13). Dans ce cas, nous avons proposé une

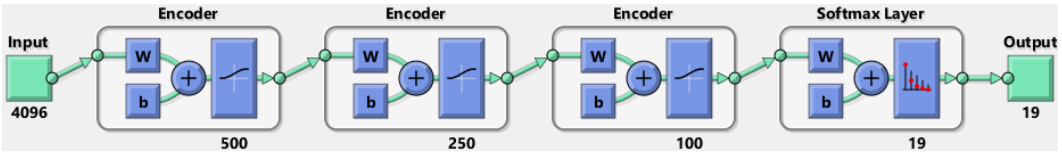


Figure 5.12 – Classification par Deep learning [132]

étape de prétraitement constitue de deux phases ; une phase pour l'extraction des silhouettes et une autre pour l'estimation de mouvement à l'aide de calcul de flux optique. L'extraction des silhouettes se fait à l'aide de l'application de modèle GMM (Gaussian Mixture Model) sur les trames extraites. Nous avons utilisé l'amplitude comme une caractéristique représentant le mouvement de l'objet. Auparavant, nous avons utilisé une technique de saut d'image dynamique rapide pour éviter les images qui contiennent des mouvements non pertinents, ce qui permet de diminuer la complexité de calcul de l'extraction de la silhouette. De plus, une nouvelle technique de représentations pour construire un concept informatif de reconnaissance de l'action humaine basée sur la superposition de silhouettes humaines est présentée. Nous avons appelé cette approche History of Binary Motion Image (HBMI). Notre méthode a été évaluée sur les bases de données Ixmas, Weizmann et KTH. Le Sparce Stacked Autoencoder (SSAE) est un exemple de stratégie d'apprentissage profond pour la détection efficace des activités humaines et une couche de sortie du type Softmax a été utilisé pour la classification.

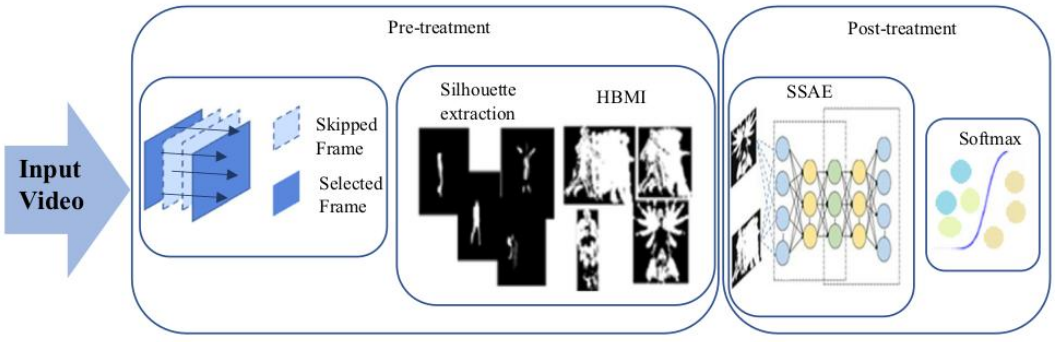


Figure 5.13 – Un système de la reconnaissance des actions humaines basé Autoencoder et HBMI

5.5 Résultats expérimentaux

5.5.1 Bases de données d'actions humaines

Afin d'évaluer l'utilisation de caractéristiques fondées sur les mouvements dérivés du flux optique à l'aide du descripteur proposé, deux ensembles de données distinctes sont envisagés pour le processus d'évaluation.

5.5.1.1 La base de données Wiezmann

La base de données Wiezmann [20, 86] est composé de 90 vidéos enregistrées avec une résolution de 180×144 à une fréquence de 25 images par seconde. La base de données contient 9 sujets différents qui sont chargés d'effectuer 10 activités élémentaires différentes. Pour notre étude, un nouveau jeu de données composé de 241 séquences vidéo est construit à partir de la base de données originale de Weizmann en annotant manuellement les vidéos pour rechercher 19 actions élémentaires différentes. Chaque séquence d'actions élémentaires s'exécute sur 15 trames qui sont toutes manuellement vérifiées et validées pour représenter une action humaine. La liste des actions élémentaires comprend : *jack_UP, jack_Down, bend_UP, bend_Down, jump_LTR, jump_RTL, run_LTR, run_RTL, side_LTR ,side_RTL, skip_LTR, skip_RTL, walk_LTR, walk_RTL, wave_One_Hand, wave_two_Hands, pjump*

5.5.1.2 La base de données UCF101

Le jeu de données UCF101 [197] contient 101 classes différentes, nous choisissons 72 vidéos pour 23 classes différentes décrivant les actions effectuées par différents utilisateurs. Les vidéos collectées à partir de ce jeu de données sont choisies de manière à éviter tout mouvement de la caméra. Dans la mesure où les mouvements de la caméra peuvent être facilement compensés à l'aide d'outils standard tels que la méthode de compensation de mouvement conventionnel (MOCO) [65] ou la compensation de mouvement en temps réel [105], cette recherche était

consacrée à la détection des actions élémentaires à partir des vidéos filmées par des caméras fixes.

5.5.2 Résultats de classification des actions humaines

Comme nous l'avons dit plus tôt, nous avons utilisé deux modes de classification ; la classification avec/sans sélection de caractéristiques. Dans les deux cas, les deux tiers de la base de données sont utilisés comme une base d'apprentissage, tandis que le tiers restant est considéré comme une base de données de test. Quand on a utilisé la procédure de la sélection des caractéristiques pour le processus de classification, nous avons obtenu une signature optimale pour les actions humaines contenant 41 caractéristiques où un taux de classification est le plus élevé. Le classifieur KNN est utilisé pour calculer le taux de classification CCR(*Correct Classification Rate*) en utilisant différentes valeurs de $k \in \{1, 3, 5\}$ et la validation croisée (*cross validation*) du type Leave-One-Out. Nous avons utilisé le classifieur KNN lors de la phase de classification en raison de sa simplicité et de la facilité de comparaison des résultats obtenus avec les autres techniques de classification appliquées sur la même base de données. Le *Cumulative Match Score* (CMS) est calculé afin d'évaluer la classification après différentes itérations ou *Rank*. Un taux de classification élevé de CCR 98,76% pour les 19 actions élémentaires a été atteint à $Rank = 1$, tandis que un taux CCR a été rapporté à 100% à $Rank = 2$. La courbe CMS est illustré dans la figure 5.14 pour le processus de classification appliqué sur la base de données Weizmann.

Le tableau 5.1 montre les résultats obtenus de la classification pour les différents classifieurs avec différentes barres spatiales appliquées sur la même base de données. Le nombre de *bars* reflète la compacité de la représentation globale des caractéristiques. Les résultats obtenus sont encourageants puisque le processus de reconnaissance est basé uniquement sur les données de mouvement du flux optique. Les caractéristiques spatiales globales sont déterminées en divisant l'image verticalement et horizontalement en un ensemble de $b \in \{5, 10, 20\}$ barres.

Pour étudier la performance des caractéristiques à l'aide de divers classifieurs, l'arbre de décision est utilisé sans aucune sélection de caractéristiques, car

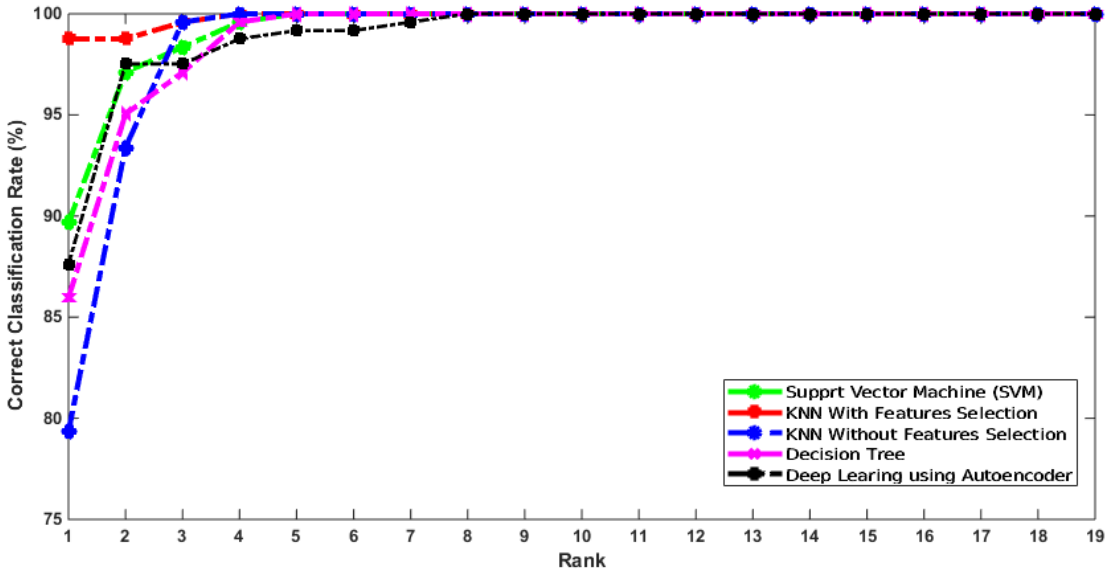


Figure 5.14 – Cumulative Match Score (CMS) à l'aide de différents classifieurs.

Table 5.1 – Effet des différentes barres spatiales sur les résultats de la classification.

Classifieurs	**Barres**		
	5	10	20
KNN			
K=1, with Features Selection	97.93	95.45	95.04
K=3, with Features Selection	**98.76**	**95.86**	**95.62**
K=5, with Features Selection	97.52	95.04	95.04
K=1, without Features Selection	83.47	85.12	81.40
K=3, without Features Selection	79.75	85.12	79.75
K=5, without Features Selection	78.51	84.71	78.51
SVM	85.95	89.67	90.08
Decision Tree	94.21	85.95	82.64

il possède sa propre méthode de sélection des caractéristiques intégrée pour les caractéristiques basées sur une entropie d'information [68, 198]. Un taux de classification de 85,95% est atteint grâce à l'utilisation d'un arbre de décision pour 10 barres et de 94,21% pour 5 barres. Une machine à vecteurs de support multi-classes (SVM) est également utilisée dans cette expérience et un taux CCR rapporté de 89,67% pour 10 barres et 90,08% pour 20 barres. On peut constater que le classifieur KNN malgré sa simplicité et avec la sélection des caractéristiques s'est

avérée utile pour obtenir des taux de reconnaissance plus élevés par rapport aux autres classifieurs. Le temps d'exécution de chaque classifieur est indiqué dans le tableau 5.2.

Table 5.2 – Temps de classification pour différents classifieurs appliqués sur la base de données Weizmann.

	Barres		
	5	10	20
KNN			
K=1	0.2497	0.2552	0.2618
K=3	0.2524	0.2574	0.2634
K=5	0.2555	0.2582	0.2646
SVM	16.5166	16.6489	16.7532
Decision Tree	0.1015	0.1017	0.1196
Deep Learning		0.2095	

5.5.3 Étude de la similarité des actions

Afin d'illustrer les résultats de la vérification visant à déduire le niveau de similarité entre deux actions humaines différentes pour toutes les paires, le ROC (Receiver Operating Characteristics) est calculé comme on le voit sur la figure 5.15. Lors de la phase de vérification, toutes les actions humaines de la base de données construite sont vérifiées séquentiellement les unes par rapport aux autres, en vérifiant si une paire donnée a la même étiquette de classe ou non. Le processus de correspondance est basé sur la distance euclidienne avec une valeur de seuil décrite dans la phase de sélection des caractéristiques pour évaluer si les deux actions humaines ont la même signification . Afin d'estimer le **FAR** (False Acceptance Rate) par rapport au **FRR** (False Rejection Rate), diverses valeurs de seuil sont introduites. En utilisant la signature d'action humaine dérivée des caractéristiques du flux optique à l'aide du descripteur basé sur l'histogramme, le système a atteint un taux d'erreur satisfaisant de 1,89 %. De plus, la correspondance des similarités est effectuée par l'analyse de la distribution entre les distances des paires appartenant aux mêmes classes par rapport aux autres classes en utilisant l'indice

de décidabilité de *Daugman* [52]. Les valeurs suivantes de l'indice de décidabilité de 0,8205 et 1,6136 sont rapportées pour la sélection des caractéristiques et les caractéristiques brutes respectivement pendant la comparaison intra-classes et inter-classes des instances. Cela montre clairement que le processus de reconnaissance de deux actions identiques basées sur la correspondance de paires est une tâche très difficile.

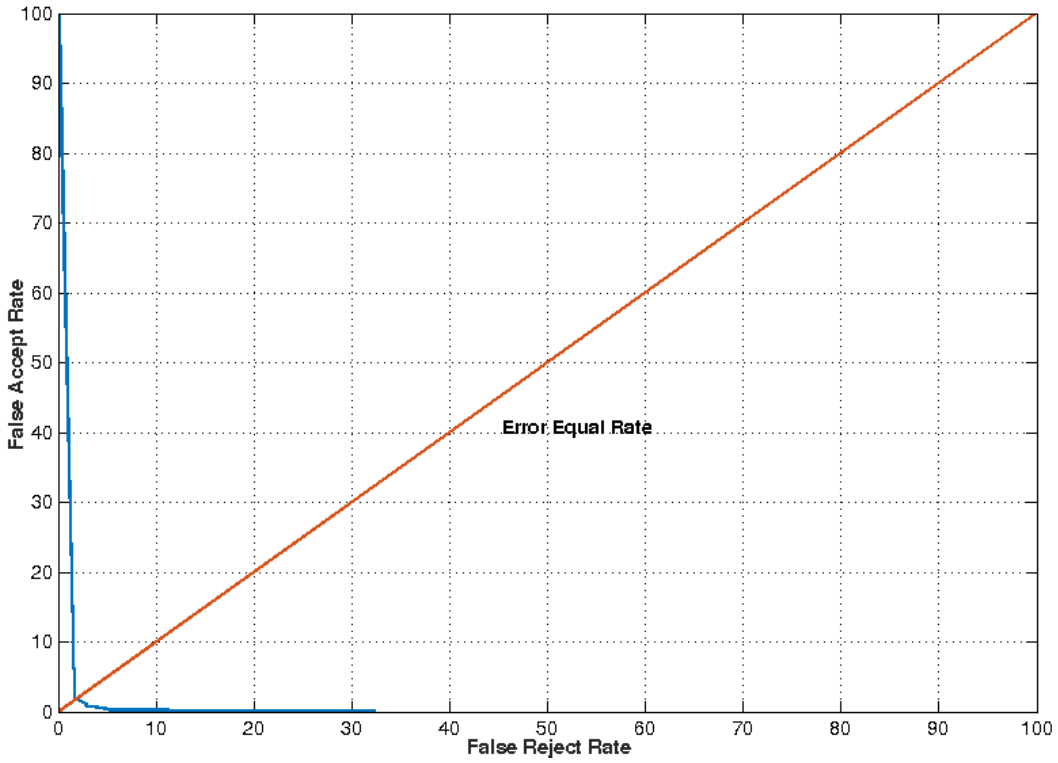


Figure 5.15 – Représentation graphique de (ROC) : Résultats de vérification de la similarité sur le jeu de données de Weizmann

La matrice de confusion indiquée à la figure 5.16visualise les résultats de vérification et de séparation entre les différentes classes d'action humaine. Les carrés blancs signifient des valeurs de séparation les plus élevées et donc une meilleure discriminabilité entre les différentes classes. La ligne diagonale foncée est la distance zéro lorsqu'on compare une classe avec elle-même. La distance euclidienne est calculée pour déduire le niveau de séparation entre deux classes comme la moyenne entre toutes les paires correspondantes. Comme toutes les caractéristiques sont déjà normalisées pendant la phase de prétraitement entre 0 et 1, on observe que certaines actions ont tendance à être presque identiques lorsqu'on utilise les caractéristiques de mouvement comme dans le cas de **Waving Hands** et **Pjump** et

autres événements. En même temps, il y a certaines similitudes entre **Running**, **Walking** et **Side Walking**.

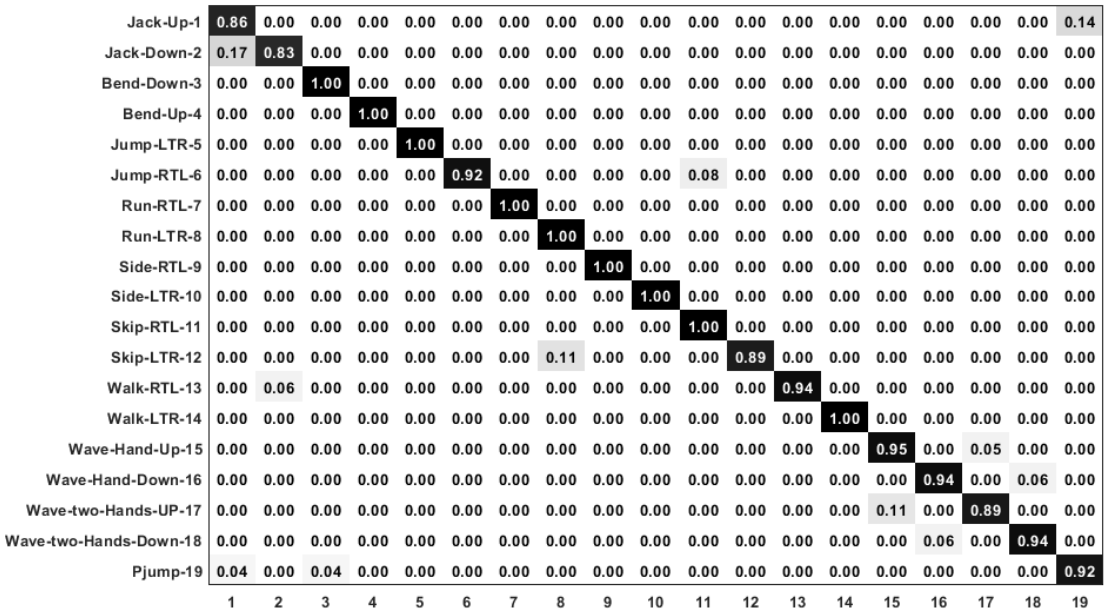


Figure 5.16 – Matrice de confusion pour la reconnaissance des actions humaines : Résultats pour l'appariement croisé de différentes classes. *Les valeurs basses reflètent une plus grande discriminabilité.*

5.5.4 Décomposition des activités complexes

En se basant sur les actions de base détectées dans le jeu de données Weizmann, d'autres expériences ont été fait pour détecter ses actions dans différents jeux de données sur des scènes plus réalistes et complexes. Nous avons annoté manuellement 1400 séquences vidéo des bases de données UCF101 [197] et KTH [190]. Ensuite, la procédure de classification est effectuée sur le jeu de données pour rechercher les actions de base à l'aide avec un seuil prédéfini qui a été défini sur la base des expériences pour la comparaison de similarité. La figure 5.17 illustre le paradigme utilisé pour la détection des actions de base à l'aide du descripteur proposé.

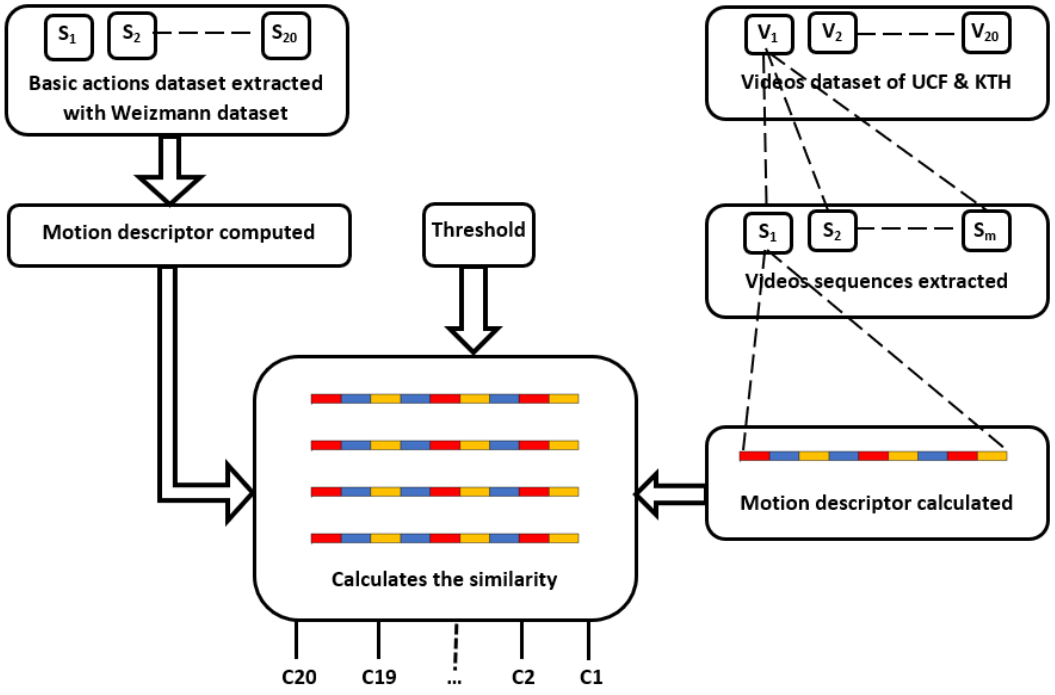


Figure 5.17 – Décomposition de la vidéo en séquences d'actions simples à l'aide du descripteur de mouvement proposé.

Le processus d'évaluation est basé sur le rappel (Recall) et la précision (Precision) en utilisant les résultats obtenus comparés aux données annotées manuellement. Les résultats sont résumés dans le tableau 5.3.

Pour définir les concepts de *rappel* et *précision*, nous devons d'abord définir les concepts suivants :

Table 5.3 – Les statistiques de la décomposition de vidéos des scènes complexes

TP	TN	FP	FN	Precision	Recall	F1Score
682	**184**	**356**	**178**	**65,70%**	**79,30%**	**71,87%**

Vrais positifs : (True Positive :TP) :Le système et l'annotateur détectent la même action.

Faux positifs : (False Positive :FP) : Le système ne détecte pas la même action étiquetée par l'annotateur.

Faux négatifs : (False Negative :FN) : Le système ne détecte pas, alors que l'annotateurs détecte une action humaine.

Vrais négatifs : (True Negative :TN) : Le système et l'annotateur ne détectent aucune action humaine dans la séquence traitée.

Les métriques estimées sont calculées comme suit :

Précision (Precision) : *Précision* ou *Spécificité* mesure la proportion de négatifs correctement identifiés comme suit :

$$Precision = \frac{TP}{TP + FP} \tag{5.10}$$

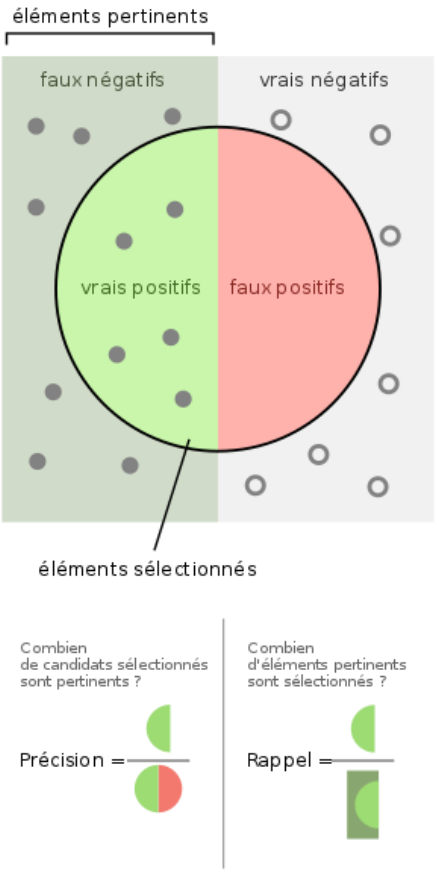


Figure 5.18 – Précision et rappel. [source : https://en.wikipedia.org/wiki/Precision_and_recall]

Rappel (Recall) : *Rappel* ou *Sensibilité* mesure la proportion de positifs correctement identifiés comme :

$$Recall = \frac{TP}{TP + FN} \tag{5.11}$$

F1 Score : *F1 Score* ou *F-measure* est calculé comme :

$$F1_{Score} = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{5.12}$$

5.5.5 Analyse comparative

L'analyse comparative de l'approche proposée par rapport aux approches existantes qui sont appliquées récemment pour la reconnaissance de l'activité humaine sur la base de données Weizmann et UCF101 est présentée dans les deux tableaux (Table 5.4 et Table 5.5) respectivement.

Table 5.4 – Résultats comparatifs pour le jeu de données Weizmann.

Methods	**CCR (%)**
Our Method : Motion Descriptors.	**98.76**
Our Method : HBMI -Deep SSAE with 10 classes.	**97.66**
Our Method : Deep Autoencoder with 10 classes.	**97.11**
Our Method : Deep Autoencoder with 19 classes.	**87.60**
Binary Motion Image and Deep Learning with 5 classes[63].	**100.00**
Multi-channel correlation filters [115].	97.80
Interest points + SIFT filters [157].	96.66
Binary motion descriptor [66].	95.81
Shape, motion and texture features [178].	94.44
Pose primitive [201].	94.40
Sequence alignment and shape context [7].	92.22
Hough Transform-Based Voting Framework [233].	92.20
Learning Mid-level Motion Features. [67].	90.50
Multiple Features [145].	90.40
Spatial-Temporal [162].	90.00

Les résultats obtenus reflètent l'efficacité de notre méthode proposée dans ce domaine difficile de la reconnaissance de l'activité humaine par la décomposition en actions de base. Dans le même but et pour enrichir le processus de comparaison, nous avons appliqué l'idée proposée par *Dobhal et al.* [63] pour la classification des actions à l'aide d'un deep learnig par la représentation des caractéristiques du mouvement humain dans la vidéo en joignant les vidéos dans une seule image

Table 5.5 – Résultats comparatifs pour le jeu de données UCF101

Methods	CCR(%)
Our Method : Motion Descriptors.	**70.00**
Bag of words [197]	44.50
Ipatio-temporal ConvNet [112]	65.40
Improved dense trajectories (IDT) [212]	85.90
IDT with higher-dimensional encodings [171]	87.90
Two-stream model (fusion by SVM) [195]	88.00
Long-term temporal convolutions [209]	92.70
TS-LSTM + Temporal-Inception [40]	94.10
Temporal Segment Networks [219]	94.20

appelée *Binary Motion Image* pour chaque action. Un taux de classification correcte de 87,60% est obtenu en utilisant Autoencoder pour 19 classes atomiques et 97,11% pour 10 classes. Ceci est dû en premier lieu au petit nombre d'éléments de chaque classe et à la difficulté de différencier entre les deux modèles de deux mouvements différents comme dans le cas des images du lever et de l'abaissement des mains.

5.6 Conclusion

Dans ce chapitre, nous avons évoqué la phase de classification de notre approche. En d'autres termes,l'interprétation des informations et la reconnaissance des activités humaines contenues dans les descripteurs étudiés au chapitre précédent. Nous avons détaillé dans ce chapitre, deux mécanismes de base pour la classification ; la classification classique à l'aide de classifieurs KNN, Arbre de décision et SVM avec ou sans la sélection des caractéristiques et la classification avancée à l'aide de Deep learning. Nous avons également expliqué un point important concernant la sélection des caractéristiques et son application en classification. En même temps, nous avons souligné les bons résultats obtenus grâce à l'introduction de ce processus comme étant une étape de prétraitement pour la classification. Dans cette optique, Nous avons choisi l'algorithme ASFFS pour la sélection des caractéristiques. Nos études empiriques et les résultats obtenus sont ensuite présentés.

Enfin, nous avons terminé ce chapitre par une étude comparative et une extension du système proposé pour la décomposition des activités complexes.

Le dernier chapitre portera sur l'analyse de performance de notre approche proposée. Plusieurs méthodes seront proposées afin de démontrer l'efficacité du système proposé.

6 Analyse de performance du système proposé

6.1 Introduction

Dans les chapitres précédents, nous avons discuté en détail les étapes à suivre pour établir le descripteur de mouvement à base de flux optique et le processus de la classification afin d'identifier les actions humaines. Certainement, pour tout travail scientifique, la phase finale de chaque système concerne la phase d'évaluation et l'étude de la performance de l'approche proposée. Cette étude, nous permet d'identifier l'adéquation entre les résultats réels que nous avons obtenus par différentes expérimentations et les objectifs précédemment identifiés. Ainsi, le rôle de l'étude de la performance d'un système est l'évaluation de celui-ci par l'introduction de plusieurs contraintes réelles ou artificielles, puis nous prenons des notes sur la réaction du système face à ces contraintes et ceci afin de connaître les aspects négatifs et positifs de l'approche proposée.

Dans notre thèse, nous proposons une méthodologie pour évaluer la performance de l'approche proposée. Celle-ci est basée sur l'introduction des restrictions ou des contraintes sur les sources de données que nous avons utilisées dans nos différentes expériences, par la suite nous observons l'impact de celles-ci sur la classification des actions. Ces restrictions devraient refléter la réalité. Dans ce contexte,

deux types de contraintes ont été retenus : la réduction de la résolution des images et les sauts ou les pertes des trames ou "Frame dropping" en Anglais. En ce qui concerne la faible résolution des images, elle reflète la vraie qualité de la vidéo enregistrée par des caméras de surveillance intégrées dans des lieux publics et dans certains lieux sensibles. En ce qui concerne l'étude des vieux films, on a remarqué souvent l'absence de certaines images de séquences vidéo, cela nous fait poser une question sur l'impact de l'absence de ces images sur la reconnaissance du contenu de la vidéo dans son ensemble. La deuxième partie de cette étude concerne l'analyse du contenu de descripteur lui-même. En d'autres termes le descripteur est composé de plusieurs parties (locale, globale et aussi spatiale et temporelle) et la question logique qui se pose est la suivante : Quel est l'effet de chacune de ces parties sur le taux de la classification ?

6.2 Analyse de performance

Pour évaluer la performance du descripteur de flux optique proposé pour la classification des actions humaines, nous avons envisagé d'explorer deux facteurs : les pertes des trames et la réduction de la résolution des images.

6.2.1 Frames dropping

Dans cette partie, nous avons étudié le problème de perte ou le saut des trames et son effet sur la reconnaissance des actions humaines. Par conséquent, cette perte est simulée par la suppression des trames progressivement, par la suite nous calculons le taux de classification pour chaque cas. Pour être plus clair, les trames sont supprimées à partir de toutes les instances de l'ensemble de données de test et la comparaison se fait avec l'ensemble de données d'entraînement original (sans perte). La figure 6.1 montre les détails de suppressions successives des trames des séquences vidéo de test.

Le tableau 6.1 indique les résultats de l'étude de la relation entre le taux de classification et le nombre de trames manquantes. Le système atteint un taux de

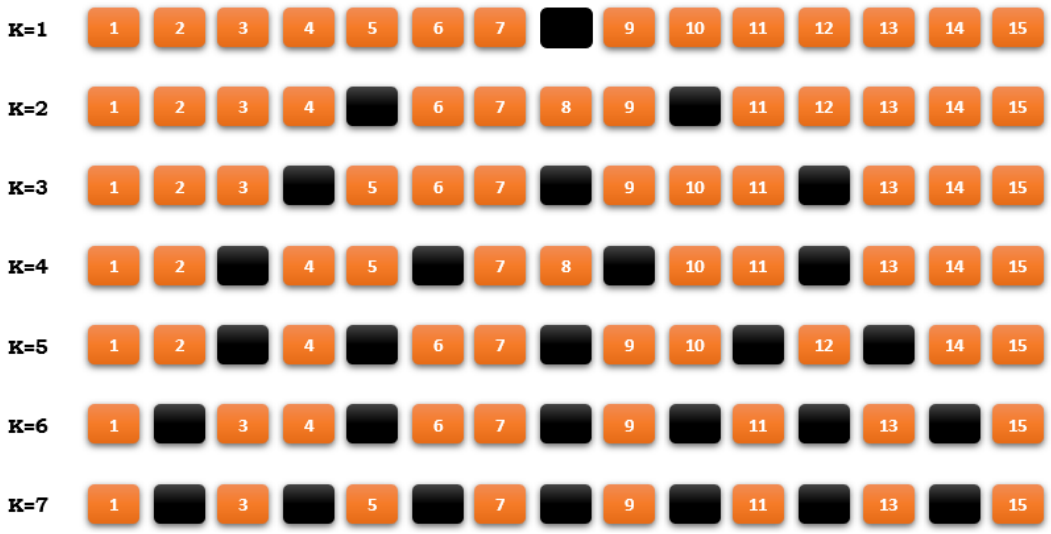


Figure 6.1 – Analyse de l'effet de perte des trames sur la reconnaissance des actions humaines (K représente le nombre de perte des trames).

réussite acceptable de 79,16% lors de la suppression d'une seule trame. Cependant, un faible taux de reconnaissance de 59,72% est signalé lors de la suppression de deux trames. ***En effet, la classification est purement basée sur la détection du mouvement à partir de la nature consécutive des trames où le saut ou l'absence des trames peut masquer ses caractéristiques vitales.***

Table 6.1 – Effet de perte des trames sur la reconnaissance des actions humaines

	Frames dropped							
	0	1	2	3	4	5	6	7
KNN								
$k = 1$	97.93	86.11	59.72	58.33	54.44	52.83	50.61	50.00
$k = 3$	98.76	79.16	59.72	57.33	44.98	44.79	44.44	44.04
$k = 5$	97.52	66.66	44.44	43.05	40.27	38.88	34.72	34.12

6.2.2 Réduction de la résolution

La deuxième contrainte qu'on a proposé pour valider la performance de notre système proposé est l'impact de la réduction de la résolution des vidéos sur le taux de la classification. Des exemples d'images de diverses résolutions sont illustrés dans la figure 6.2. Pour cette raison, nous réduisons la taille de la trame pour

toutes les données de 90% à 50% avec des diminutions de 10% tandis que le taux de classification (CCR) est calculé séparément pour chaque nouvelle résolution. Sachant que dans la technologie de la surveillance, la résolution des images est toujours faible.

(a) 70% :101x126 (b) 60% :87x108 (c) 50% :72x90

Figure 6.2 – Différentes résolutions pour le jeu de données Weizmann.

Le tableau 6.2 présente deux résultats expérimentaux. Dans le premier, le sous-ensemble original des caractéristiques est utilisé pour les résolutions réduites. Le système montre un taux de reconnaissance acceptable de 80% pouvant être obtenu même dans le cas de 116 x144 .

Table 6.2 – Effet de la réduction de la résolution sur la reconnaissance de l'action humaine

KNN	**Resolution**					
	100%	90%	80%	70%	60%	50%
	144x180	130x162	116x144	101x126	87x108	72x90
Without Features Selection						
$k = 1$	85.12	84.30	80.16	73.96	71.48	71.07
$k = 3$	85.12	79.75	79.34	72.72	69.83	67.76
$k = 5$	84.71	77.68	76.03	72.72	69.01	66.94
With Features Selection						
$k = 1$	97.93	93.80	90.08	89.67	83.47	77.27
$k = 3$	98.76	89.66	89.25	88.84	82.23	74.38
$k = 5$	97.52	88.84	88.42	86.36	80.57	72.72

Selon ce tableau, nous pouvons déduire la même observation où le processus de la sélection des caractéristiques est appliqué. Cela montre la puissance de la

méthode proposée surtout pour les systèmes de surveillance qui reposent principalement sur des caméras avec des résolutions réduites. Il est également clair que les résultats obtenus en utilisant la sélection des caractéristiques sont bien meilleurs que leurs similaires dans le premier cas. Dans le dernier cas, on peut dire que : ***Le processus de classification n'est pas trop affecté par la résolution faible des images. Nous pouvons améliorer ces résultats par leur renforcement par le processus de la sélection des caractéristiques.***

6.3 Analyse des caractéristiques

Une analyse expérimentale est effectuée afin d'étudier la distribution des caractéristiques du flux optique et de déterminer quels sont les indicateurs de mouvement qui jouent un rôle clé dans la reconnaissance des activités humaines. Les composantes de la signature de l'action humaine dans l'histogramme sont évaluées séparément pour déterminer leur contribution et leur puissance de reconnaissance au cours du processus de classification. Nous avons trouvé 17293 sous-ensembles de caractéristiques dans le cadre de cette étude empirique de telle sorte que chaque sous-ensemble atteint un taux de classification de 98,76% selon le classifieur KNN. Tous les sous-ensembles de caractéristiques ont des tailles de caractéristiques allant de 28 à 100. Le fait de choisir un grand nombre de sous-ensembles permettrait d'obtenir des résultats objectifs et précis pour l'analyse. Les distributions et les résultats de la classification des actions humaines de différents types de caractéristiques sont présentés au tableau 6.3.

Table 6.3 – Analyse des caractéristiques pour déterminer la contribution de chaque type dans la reconnaissance des actions humaine

Caractéristiques	**Distribution (%)**	**CCR (%)**
Caractéristiques locales	00.08	88.42
Caractéristiques globales-*Temporelles*	90.58	93.39
Caractéristiques globales-*Spatiales*	09.34	65.29

Les résultats de la distribution montrent clairement le type de la caractéristique dominante, mais ne permettent pas de mesurer son caractère potentiellement

discriminatoire. Alternativement, la pertinence de la reconnaissance des caractéristiques de flux optique est approchée par l'utilisation du taux de classification (CCR). Les résultats obtenus montrent l'importance des caractéristiques temporelles globales qui y contribuent avec près de 90% pour les vecteurs de caractéristiques tout en atteignant un taux de reconnaissance de 93%. Dans le cas des caractéristiques locales qui décrivent les vecteurs de flux optiques sans information temporelle ou spatiale, une contribution marginale de 0,08% est rapportée avec un taux de classification de 88%. ***La combinaison des caractéristiques locales et globales montre une influence considérable pour augmenter le taux de classification pour la reconnaissance de l'action humaine.***

6.4 Conclusion

Dans ce chapitre, nous avons mené une étude de performance de notre approche proposée pour la reconnaissance des activités humaines. Nous avons également observé l'impact limité de la réduction de la résolution des images sur la classification des actions humaines, lorsque la sélection des caractéristiques est utilisée, son effet apparaît à peine, cela indique la rigidité de l'approche proposée. Dans le même contexte, nous avons étudié l'effet de la suppression des trames de séquences vidéo sur les résultats de la classification et les résultats étaient attendus ; le taux de classification à diminuer très rapidement, ce qui indique que la reconnaissance du mouvement est directement liée à la séquence des trames.

Nous avons fini cette étude par une analyse exploratoire afin d'étudier l'impact de la distribution des caractéristiques du flux optique et de déterminer quels sont les indices de mouvement qui jouent un rôle crucial dans la reconnaissance de l'activité humaine.

7 Conclusions et perspectives

7.1 Conclusions

À travers ce manuscrit, nous avons présenté un ensemble d'outils permettant de résoudre certains problèmes de la reconnaissance des activités humaines à partir des séquences vidéo. Les motivations liées à ces travaux sont l'extraction, la représentation et la reconnaissance automatique des actions et des activités humaines. Le déploiement des méthodes automatiques de vision par ordinateur pour la reconnaissance des activités humaines a une importance capitale pour de nombreuses applications comme la surveillance visuelle automatique, l'analyse sportive, l'interaction homme-machine,...

Dans cette étude, un descripteur est introduit pour la représentation des caractéristiques visuelle des mouvements basés sur le flux optique appliqué sur un ensemble de trames consécutives pour la classification des actions humaines. Un histogramme des caractéristiques de mouvement est produit en tenant compte des caractéristiques locales et globales intégrées dans le flux optique. La sélection des caractéristiques est effectuée pour obtenir les caractéristiques les plus discriminantes. Afin d'évaluer le descripteur proposé pour la reconnaissance des activités humaines, des tests ont été fait sur deux bases de données publiques Weizmann et UCF101, les potentiels de l'approche proposée ont été confirmés avec un taux élevé de classification de 98,76% et 70% respectivement, de reconnaissance des

actions humaines élémentaires. Les résultats obtenus sont en accord avec les premières études psychologiques indiquant que le mouvement humain est suffisant pour la perception des activités humaines. D'autres évaluations empiriques sont effectuées afin de déterminer les performances du descripteur introduit pour traiter les différentes résolutions et les sauts des trames.

7.2 Perspectives

Les différents travaux développés dans cette thèse ont permis de découvrir des nouvelles pistes de recherche sur la reconnaissance automatique des activités humaines. Ces pistes sont à la fois des perspectives d'amélioration de notre méthode, mais aussi des perspectives d'application de cette méthode à d'autres domaines de recherche.

7.2.1 Perspectives relatives au modèle proposé

— **Effet du nombre des secteurs :** le modèle du descripteur proposé pour la discrétisation de l'orientation de mouvement a été calculé seulement sur 8 secteurs. Nous pensons qu'il est important d'enrichir cette recherche pour voir l'impact du nombre de secteurs sur la qualité du descripteur.

— **Action VS activité** : L'objectif initial de notre descripteur est dédié à la reconnaissance des actions humaines. Nous avons tenté d'évoluer ce descripteur afin qu'il soit apte à connaître des activités complexes. Dans ce contexte, nous avons proposé une méthodologie pour la décomposition des activités complexes. Nous avons trouvé des résultats raisonnables, mais nous pensons qu'il faudrait ajouter d'autres améliorations.

— **Problème du mouvement de la caméra :** parmi les avantages de notre approche proposée, nous n'avons pas besoin de l'extraction du fond de l'image mais ce n'est pas toujours le cas, en particulier quand la caméra est en mouvement. Récemment, des grands progrès ont été accomplis dans le domaine de détection des poses. Cela nous permet de nous concentrer sur

la personne en déplacement sans faire attention au fond de l'image ni au mouvement de la caméra.

— **Applications en temps réelles :** Les algorithmes utilisés pour l'estimation de flux optique et la sélection des caractéristiques consomme beaucoup de temps et leurs complexités dépendent exponentiellement de la taille de la base de données utilisées. Cela nécessite de penser à d'autres algorithmes plus efficaces et moins gourmands en temps machine.

7.2.2 Pistes de recherche ouvertes par la méthode

À l'issue de ce manuscrit, plusieurs points s'avèrent intéressants à explorer et à développer. Nous avons identifié en particulier les points suivants qui nous paraissent les plus intéressants :

— **Extension du modèle de la décomposition des activités humaines :** Nous avons proposé un modèle de décomposition des activités en séquences d'actions simples. En revanche, nous proposons des recherches sur la définition d'un modèle représentant une activité à base d'un dictionnaire des actions simples où des mots qui nous permet de connaître ces activités à partir des séquences vidéo grâce à l'utilisation de la notion de similarité entre les actions et d'un classifieur du type Bag of Visual Words (BOVW) [47].

— **Deep learning et la reconnaissance des activités :** dans cette thèse, nous avons proposé deux approches pour la reconnaissance des activités humaines à base de classifieurs du type Deep learning. Les deux approches partagent la propriété suivante : les données en entrées sont des matrices 2Ds des caractéristiques. Il existe d'autres modèles pour représenter les caractéristiques en 3D ou avec des modèles hybrides qui peuvent être utilisés pour étudier leur efficacité. D'autre part, il existe d'autres structures et architectures pour les réseaux de neurones du type deep learning qui doit également être testé [213].

— **Détection anormale des activités et notre descripteur de mouvement** : l'une des tâches les plus difficiles en vision par ordinateur est l'analyse

de l'activité humaine dans des scènes encombrées. Bien que la compréhension des actions accomplies par les individus soit un problème à résoudre, l'analyse des scènes de foule fait face à des défis encore plus grands. Nous proposons notre descripteur à base de l'estimation de flux optique (la direction + la vitesse) comme une solution pour la détection des activités anormales d'une foule dans une séquence vidéo. Des travaux de recherche ont été entrepris pour détecter les activités anormales dans ce sens et parmi eux on cite :[177, 36].

— **Prédiction des activités :** La prédiction de l'activité humaine est un processus probabiliste qui consiste à déduire les activités en cours à partir de vidéos ne contenant que le début des activités. L'objectif est de permettre la reconnaissance précoce des activités inachevées plutôt que la classification a posteriori des activités terminées. Les méthodologies de prévision d'activité sont particulièrement nécessaires pour les systèmes de surveillance qui sont nécessaires pour prévenir les crimes et les activités dangereuses [188, 182]. Nous proposons d'utiliser le mécanisme qu'on a proposé pour la décomposition des activités en séquences des actions élémentaires ou les mots visuels au sens de Bow (Bag of Words) pour résoudre ce problème. Cela signifie qu'il faut étudier la similarité entre la vidéo en entrée (incomplète) avec les bases des activités complètes.

— **Estimation de pose et reconnaissance des activités :** Récemment, le domaine de l'estimation de pose a évolué de façon remarquable, notamment par l'introduction de deep learning [203, 160, 150]. En revanche, pour minimiser le temps de calcul de l'estimation de flux optique, nous proposons d'utiliser l'estimation de pose comme étant une étape de prétraitement pour un système de la reconnaissance des activités humaines. Dans le même contexte, il existe des études qui montrent la possibilité de la détection des poses d'une foule des personnes [32, 33], cela nous donne une plus grande chance de la reconnaissance des activités de plusieurs personnes en même temps.

Les travaux présentés dans ce manuscrit constituent un premier pas vers des approches pour la reconnaissance des activités humaines individuellement ou en groupes. Nous espérons qu'ils ouvriront la voie à de nouveaux travaux qui lèvent les obstacles restants.

Bibliographie

[1] Jake K Aggarwal and Michael S Ryoo. Human activity analysis : A review. *ACM Computing Surveys (CSUR)*, 43(3) :16, 2011.

[2] Mohiuddin Ahmad and Seong-Whan Lee. Human action recognition using shape and clg-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7) :2237–2252, 2008.

[3] Kelson RT Aires, Andre M Santana, and Adelardo AD Medeiros. Optical flow using color information : preliminary results. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1607–1611. ACM, 2008.

[4] Koichiro Akita. Image sequence analysis of real world human motion. *Pattern recognition*, 17(1) :73–83, 1984.

[5] Analí Alfaro, Domingo Mery, and Alvaro Soto. Human action recognition from inter-temporal dictionaries of key-sequences. In *Pacific-Rim Symposium on Image and Video Technology*, pages 419–430. Springer, 2013.

[6] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 32(2) :288–303, 2010.

[7] Sultan Almotairi and Eraldo Ribeiro. Action classification using sequence alignment and shape context. In *The Twenty-Seventh International Flairs Conference*, 2014.

[8] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3) :175–185, 1992.

[9] Maryam Asadi-Aghbolaghi, Albert Clapés, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. A survey on deep learning based approaches for action and gesture recognition in image sequences. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 476–483. IEEE, 2017.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014.

[11] UABUA Bakar, Hemant Ghayvat, SF Hasanm, and SC Mukhopadhyay. Activity and anomaly detection in smart home : A survey. In *Next Generation Sensors and Systems*, pages 191–220. Springer, 2016.

[12] John L Barron, David J Fleet, Steven S Beauchemin, and TA Burkitt. Performance of optical flow techniques. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 236–242. IEEE, 1992.

[13] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3) :433–466, 1995.

[14] Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. Towards biologically plausible deep learning. *arXiv preprint arXiv :1502.04156*, 2015.

[15] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[16] Douglas A Bernstein. Essentials of psychology. cengage learning, 123–124. Technical report, ISBN 978-0-495-90693-3, 2010.

[17] Christel Bidet-Ildei, Jean-Pierre Orliaguet, and Yann Coello. Rôle des représentations motrices dans la perception visuelle des mouvements humains. *LAnnee psychologique*, 111(2) :409–445, 2011.

[18] Geoffrey P Bingham, Richard C Schmidt, and Lawrence D Rosenblum. Dynamics and the orientation of kinematic forms in visual event recognition. *Journal of Experimental Psychology : Human Perception and Performance*, 21(6) :1473, 1995.

[19] Gary Bishop, Greg Welch, et al. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, 8(27599-3175) :59, 2001.

[20] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[21] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *null*, pages 1395–1402. IEEE, 2005.

[22] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3) :257–267, 2001.

[23] I. Bouchrika, J. N. Carter, and M. S. Nixon. Recognizing people in non-intersecting camera views. In *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, pages 1–6, Dec 2009.

[24] Imed Bouchrika. *Gait analysis and recognition for automated visual surveillance*. PhD thesis, University of Southampton, 2008.

[25] Imed Bouchrika. Evidence evaluation of gait biometrics for forensic investigation. In *Multimedia Forensics and Security*, pages 307–326. Springer, 2017.

[26] Imed Bouchrika, John N Carter, Mark S Nixon, R Morzinger, and Georg Thallinger. Using gait features for improving walking people detection. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3097–3100. IEEE, 2010.

[27] Imed Bouchrika and Mark S Nixon. Exploratory factor analysis of gait recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[28] William Brendel and Sinisa Todorovic. Video object segmentation by tracking regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 833–840. IEEE, 2009.

[29] C. Brown and C. Brown. *Advances in Computer Vision*. Number vol. 1. Taylor & Francis, 2014.

[30] Andrew Burton and John Radford. *Thinking in perspective : critical essays in the study of thought processes*, volume 646. Routledge, 1978.

[31] Meng Cai, Yongzhe Shi, and Jia Liu. Deep maxout neural networks for speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 291–296. IEEE, 2013.

[32] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose : realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv :1812.08008*, 2018.

[33] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[34] Claudette Cedras and Mubarak Shah. Motion-based recognition : A survey. *Image and Vision Computing*, 13(2) :129, 1995.

[35] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15) :1799–1807, 2013.

[36] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection : A survey. *arXiv preprint arXiv :1901.03407*, 2019.

[37] Haiyang Chao, Yu Gu, and Marcello Napolitano. A survey of optical flow techniques for robotics navigation applications. *Journal of Intelligent & Robotic Systems*, 73(1-4) :361–372, 2014.

[38] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6) :633–659, 2013.

[39] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE, 2009.

[40] MH Chen, Z Kira, et al. Ts-lstm and temporal-inception : Exploiting spatiotemporal dynamics for activity recognition [j]. *arXiv preprint arXiv*, 1703 :10667, 2017.

[41] Guangchun Cheng, Yiwen Wan, Abdullah N Saudagar, Kamesh Namuduri, and Bill P Buckles. Advances in human action recognition : A survey. *arXiv preprint arXiv :1501.05964*, 2015.

[42] Hassan Chouaib. Sélection de caractéristiques : méthodes et applications. *Paris Descartes University : Paris, France*, 2011.

[43] Dan CireşAn, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural networks*, 32 :333–338, 2012.

[44] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3) :673–682, 2017.

[45] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

[46] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[47] Xavier Cortés, Donatello Conte, and Hubert Cardot. A new bag of visual words encoding method for human action recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2480–2485. IEEE, 2018.

[48] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 2003.

[49] James E Cutting, Cassandra Moore, and Roger Morrison. Masking the motions of human gait. *Perception & psychophysics*, 44(4) :339–347, 1988.

[50] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[51] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(3) :131–156, 1997.

[52] John Daugman. How iris recognition works. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1) :21–30, 2004.

[53] César Roberto De Souza, Adrien Gaidon, Eleonora Vig, and Antonio Manuel López. Sympathy for the details : Dense trajectories and hybrid classification architectures for action recognition. In *European Conference on Computer Vision*, pages 697–716. Springer, 2016.

[54] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.

[55] L. Deng and D. Yu. *Deep Learning : Methods and Applications*. now, 2014.

[56] Li Deng, Dong Yu, et al. Deep learning : methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4) :197–387, 2014.

[57] Andrew M Derrington, Harriet A Allen, and Louise S Delicato. Visual mechanisms of motion analysis and motion perception. *Annu. Rev. Psychol.*, 55 :181–205, 2004.

[58] Simon-Frederic Desage. *Contraintes et opportunités pour l'automatisation de l'inspection visuelle au regard du processus humain*. PhD thesis, Grenoble Alpes, 2015.

[59] Romuald Deshayes, Tom Mens, and Philippe Palanque. A generic framework for executable gestural interaction models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, pages 35–38. IEEE, 2013.

[60] Pravin Dhulekar, ST Gandhe, Harshada Chitte, and Komal Pardeshi. Human action recognition : An overview. In *Proceedings of the International Conference on Data Engineering and Communication Technology*, pages 481–488. Springer, 2017.

[61] TG Ditterrich. Machine learning research : four current direction. *Artificial Intelligence Magzine*, 4 :97–136, 1997.

[62] Winand H Dittrich. Action categories and the perception of biological motion. *Perception*, 22(1) :15–22, 1993.

[63] Tushar Dobhal, Vivswan Shitole, Gabriel Thomas, and Girisha Navada. Human activity recognition using binary motion image and deep learning. *Procedia Computer Science*, 58 :178–185, 2015.

[64] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf : A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[65] BangKui Fan, ZeGang Ding, WenBin Gao, and Teng Long. An improved motion compensation method for high resolution uav sar imaging. *Science China Information Sciences*, 57(12) :1–13, 2014.

[66] Abassin Sourou Fangbemi, Bin Liu, Nenghai Yu, and Yanxiang Zhang. Binary proximity patches motion descriptor for action recognition in videos. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*, page 17. ACM, 2018.

[67] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[68] Usama M Fayyad and Keki B Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine learning*, 8(1) :87–102, 1992.

[69] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[70] Yinfu Feng, Mingming Ji, Jun Xiao, Xiaosong Yang, Jian J Zhang, Yueting Zhuang, and Xuelong Li. Mining spatial-temporal patterns and structural sparsity for human motion data denoising. *IEEE transactions on cybernetics*, 45(12) :2693–2706, 2015.

[71] Basura Fernando, Peter Anderson, Marcus Hutter, and Stephen Gould. Discriminative hierarchical rank pooling for activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1924–1932, 2016.

[72] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015.

[73] Basura Fernando and Stephen Gould. Learning end-to-end video classification with rank-pooling. In *International Conference on Machine Learning*, pages 1187–1196, 2016.

[74] Borko Furht, Joshua Greenberg, and Raymond Westwater. *Motion estimation algorithms for video compression*, volume 379. Springer Science & Business Media, 2012.

[75] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Activity representation with motion hierarchies. *International journal of computer vision*, 107(3) :219–238, 2014.

[76] Dariu M Gavrila, Larry S Davis, et al. Towards 3-d model-based tracking and recognition of human movement : a multi-view approach. In *International workshop on automatic face-and gesture-recognition*, pages 272–277. Citeseer, 1995.

[77] Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Evans Etrue, and Giovanni Zappella. On context-dependent clustering of bandits. *arXiv preprint arXiv :1608.03544*, 2016.

[78] Khalid Ghoul, Mohamed Berkane, and Mohamed Chawki Batouche. Phase correlation method to the optical flow using neuronal networks. In *Multimedia Computing and Systems (ICMCS), 2016 5th International Conference on*, pages 71–75. IEEE, 2016.

[79] J.J. Gibson. *The perception of the visual world.* Houghton Mifflin, 1950.

[80] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

[81] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. Dense image registration through mrfs and efficient linear programming. *Medical image analysis*, 12(6) :731–741, 2008.

[82] Mariem Gnouma, Ammar Ladjailia, Ridha Ejbali, and Mourad Zaied. Stacked sparse autoencoder and history of binary motion image for human activity recognition. *Multimedia Tools and Applications*, 78(2) :2157–2179, 2019.

[83] Nigel H Goddard. The perception of articulated motion : recognizing moving light displays. Technical report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE, 1992.

[84] E Bruce Goldstein and James Brockmole. *Sensation and perception.* Cengage Learning, 2016.

[85] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[86] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12) :2247–2253, December 2007.

[87] Frédéric Grandidier. *Un nouvel algorithme de sélection de caractéristiques : application à la lecture automatique de l'écriture manuscrite*. PhD thesis, École de technologie supérieure, 2003.

[88] Richard C. Schmidt Gregory A. Burton. *Studies in perception and action VI : Eleventh International Conference on Perception and Action : June 24-29, 2001, Storrs, CT, USA*. L. Erlbaum Associates, 1 edition, 2001.

[89] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C : Signal Processing (Cat. No.94CH3440-5)*, volume 2, pages 325–329 vol.2, Oct 1994.

[90] M.A. Hall. *Correlation-based Feature Selection for Machine Learning*. University of Waikato, 1999.

[91] A Hein, DE Goldberg, and DP Michelfelder. Identification and bridging of semantic gaps in the context of multi-domain engineering. In *Forum on Philosophy, Engineering & Technology*, 2010.

[92] David Held, Sebastian Thrun, and Silvio Savarese. Deep learning for single-view instance recognition. *arXiv preprint arXiv :1507.08286*, 2015.

[93] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8) :1735–1780, 1997.

[94] Berthold Horn, Berthold Klaus, and Paul Horn. *Robot vision*. MIT press, 1986.

[95] Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics, 1981.

[96] Thanarat Horprasert, David Harwood, and Larry S Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE ICCV*, volume 99, pages 1–19. Citeseer, 1999.

[97] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2) :179–187, 1962.

[98] Cheng-Ming Huang, Yi-Ru Chen, and Li-Chen Fu. Real-time object detection and tracking on a moving camera platform. In *ICCAS-SICE, 2009*, pages 717–722. IEEE, 2009.

[99] Glyn W Humphreys and Vicki Bruce. *Visual cognition : Computational, experimental and neuropsychological perspectives*. Psychology Press, 1989.

[100] Nazli Ikizler, Ramazan Gokberk Cinbis, and Pinar Duygulu Sahin. Human action recognition with line and flow histograms. *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[101] Michael Isard and Andrew Blake. Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29 :5–28, 1998.

[102] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3) :194, 2001.

[103] Mihir Jain, Herve Jegou, and Patrick Bouthemy. Better exploiting motion for better action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2555–2562, 2013.

[104] Ahmad Jalal, Maria Mahmood, and Abdul S Hasan. Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments. In *IEEE International Conference on Applied Sciences and Technology*, 2019.

[105] Klaus Janschek, Valerij Tchernykh, and Serguei Dyblenko. Integrated camera motion compensation by real-time image motion tracking and image deconvolution. In *Advanced Intelligent Mechatronics. Proceedings, 2005 IEEE/ASME International Conference on*, pages 1437–1444. IEEE, 2005.

[106] Yu-Gang Jiang, Qi Dai, Wei Liu, Xiangyang Xue, and Chong-Wah Ngo. Human action recognition in unconstrained videos by explicit motion modeling. *IEEE Transactions on Image Processing*, 24(11) :3781–3795, 2015.

[107] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2) :201–211, 1973.

[108] Vadim Kantorov and Ivan Laptev. Efficient feature extraction, encoding and classification for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2593–2600, 2014.

[109] Purushottam Kar, Shuai Li, Harikrishna Narasimhan, Sanjay Chawla, and Fabrizio Sebastiani. Online optimization methods for the quantification problem. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1625–1634. ACM, 2016.

[110] IA Karaulova, Peter M Hall, and A David Marshall. A hierarchical model of dynamics for tracking people with a single video camera. In *BMVC*, pages 1–10, 2000.

[111] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[112] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[113] Shian-Ru Ke, Hoang Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. A review on video-based human activity recognition. *Computers*, 2(2) :88–131, 2013.

[114] Yan Ke, Rahul Sukthankar, and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[115] Hamed Kiani, Terence Sim, and Simon Lucey. Multi-channel correlation filters for human action recognition. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1485–1489. IEEE, 2014.

[116] Kye Kyung Kim, Soo Hyun Cho, Hae Jin Kim, and Jae Yeon Lee. Detecting and tracking moving object using an active camera. In *Advanced Communication Technology, 2005, ICACT 2005. The 7th International Conference on*, volume 2, pages 817–820. IEEE, 2005.

[117] Bernd Kitt, Benjamin Ranft, and Henning Lategahn. Block-matching based optical flow estimation with reduced search space based on geometric constraints. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1104–1109. IEEE, 2010.

[118] Orit Kliper-Gross, Yaron Gurovich, Tal Hassner, and Lior Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision*, pages 256–269. Springer, 2012.

[119] Teddy Ko. A survey on behavior analysis in video surveillance for homeland security applications. *2008 37th IEEE Applied Imagery Pattern Recognition Workshop*, pages 1–8, 2008.

[120] Yu Kong and Yun Fu. Bilinear heterogeneous information machine for rgb-d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1054–1062, 2015.

[121] Yu Kong and Yun Fu. Max-margin action prediction machine. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 1844–1858, 2016.

[122] Yu Kong and Yun Fu. Max-margin heterogeneous information machine for rgb-d action recognition. *International Journal of Computer Vision*, 123(3) :350–371, 2017.

[123] Yu Kong and Yun Fu. Human action recognition and prediction : A survey. *CoRR*, abs/1806.11230, 2018.

[124] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1481, 2017.

[125] Lynn T Kozlowski and James E Cutting. Recognizing the gender of walkers from point-lights mounted on ankles : Some second thoughts. *Attention, Perception, & Psychophysics*, 23(5) :459–459, 1978.

[126] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[127] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB : a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[128] Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller, and Sebastian Möller. I'm home : Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies*, 69(11) :693–704, 2011.

[129] Ammar Ladjailia, Imed Bouchrika, Nouzha Harrati, and Zohra Mahfouf. Encoding human motion for automated activity recognition in surveillance applications. In *Computer Vision : Concepts, Methodologies, Tools, and Applications*, pages 2042–2064. IGI Global, 2018.

[130] Ammar Ladjailia, Imed Bouchrika, Hayet Farida Merouani, and Nouzha Harrati. Automated detection of similar human actions using motion descriptors. In *2015 16th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, pages 398–403. IEEE, 2015.

[131] Ammar Ladjailia, Imed Bouchrika, Hayet Farida Merouani, and Nouzha Harrati. On the use of local motion information for human action recognition via feature selection. In *2015 4th International Conference on Electrical Engineering (ICEE)*, pages 1–4. IEEE, 2015.

[132] Ammar Ladjailia, Imed Bouchrika, Hayet Farida Merouani, Nouzha Harrati, and Zohra Mahfouf. Human activity recognition via optical flow : decomposing activities into basic actions. *Neural Computing and Applications*, pages 1–14, 2019.

[133] Zhengzhong Lan, Ming Lin, Xuanchong Li, Alex G Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid : Multi-skip feature stacking for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 204–212, 2015.

[134] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3) :107–123, 2005.

[135] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE*, 15(3) :1192–1209, 2013.

[136] Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[137] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553) :436, 2015.

[138] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5) :421–436, 2018.

[139] Kang Li and Yun Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE transactions on pattern analysis and machine intelligence*, 36(8) :1644–1657, 2014.

[140] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548. ACM, 2016.

[141] Yingwei Li, Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Vlad3 : Encoding dynamics of deep features for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1951–1960, 2016.

[142] A. Lieury. *Manuel de psychologie générale*. Psycho sup. Dunod, 1997.

[143] Wei Ying Lim, Angela Ong, Lay Lian Soh, and Adam Sufi. Teachers' voices and change : The structure and agency dialectics that shaped teachers' pedagogy toward deep learning. In *Future Learning in Primary Schools*, pages 147–158. Springer, 2016.

[144] James J Little and Alessandro Verri. Analysis of differential and matching methods for optical flow. In *Visual Motion, 1989., Proceedings. Workshop on*, pages 173–180. IEEE, 1989.

[145] Jingen Liu, Saad Ali, and Mubarak Shah. Recognizing human actions using multiple features. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[146] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE, 1999.

[147] Wei-Lwun Lu and James J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pages 6–6, 2006.

[148] Xin Lu, Qiong Liu, and Shunichiro Oe. Recognizing non-rigid human actions using joints tracking in space-time. *International Conference on Information Technology : Coding and Computing, 2004. Proceedings. ITCC 2004.*, 1 :620–624 Vol.1, 2004.

[149] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[150] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[151] Zohra Mahfouf, Hayet Farida Merouani, Imed Bouchrika, and Nouzha Harrati. Investigating the use of motion-based features from optical flow for gait recognition. *Neurocomputing*, 283 :140–149, 2018.

[152] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10 :94, 2016.

[153] Fabio Martínez, Antoine Manzanera, and Eduardo Romero. A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance. In *Multimedia and Signal Processing*, pages 267–274. Springer, 2012.

[154] Roberto Melfi, Shripad Kondra, and Alfredo Petrosino. Human activity modeling by spatio temporal textural appearance. *Pattern Recognition Letters*, 34(15) :1990–1994, 2013.

[155] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Unsupervised learning using sequential verification for action recognition. *ArXiv*, abs/1603.08561, 2016.

[156] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2) :90–126, 2006.

[157] Mona M Moussa, Elsayed Hamayed, Magda B Fayek, and Heba A El Nemr. An enhanced method for human action recognition. *Journal of advanced research*, 6(2) :163–169, 2015.

[158] Don Murray and Anup Basu. Motion tracking with an active camera. *IEEE transactions on pattern analysis and machine intelligence*, 16(5) :449–459, 1994.

[159] Randal C Nelson and Ramprasad Polana. Qualitative recognition of motion using temporal texture. *CVGIP : Image understanding*, 56(1) :78–89, 1992.

[160] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[161] Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona. Human detection from images and videos : A survey. *Pattern Recognition*, 51 :148–175, 2016.

[162] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3) :299–318, 2008.

[163] Seyed Yahya Nikouei, Yu Chen, Sejun Song, Ronghua Xu, Baek-Young Choi, and Timothy R Faughnan. Real-time human detection as an edge service enabled by a lightweight cnn. In *2018 IEEE International Conference on Edge Computing (EDGE)*, pages 125–129. IEEE, 2018.

[164] Abhijit S Ogale, Alap Karapurkar, and Yiannis Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *Dynamical vision*, pages 115–126. Springer, 2007.

[165] Godwin Ogbuabor and Robert La. Human activity recognition for healthcare using smartphones. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pages 41–46. ACM, 2018.

[166] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(3) :710–719, 2005.

[167] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583) :607, 1996.

[168] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Proceedings of the IEEE international conference on computer vision*, pages 1817–1824, 2013.

[169] Olusegun Oshin, Andrew Gilbert, and Richard Bowden. Capturing relative motion and finding modes for action recognition in the wild. *Computer Vision and Image Understanding*, 125 :155–171, 2014.

[170] Josh Patterson and Adam Gibson. *Deep Learning : A Practitioner's Approach*. O'Reilly, Beijing, 2017.

[171] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition : Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150 :109–125, 2016.

[172] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *European Conference on Computer Vision*, pages 581–595. Springer, 2014.

[173] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4) :695–706, 2006.

[174] Ramprasad Polana and Randal Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, pages 77–82. IEEE, 1994.

[175] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6) :976–990, 2010.

[176] Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11) :1119–1125, 1994.

[177] Hamidreza Rabiee, Hossein Mousavi, Moin Nabi, and Mahdyar Ravanbakhsh. Detection and localization of crowd behavior using a novel tracklet-based model. *International Journal of Machine Learning and Cybernetics*, 9(12) :1999–2010, 2018.

[178] Saimunur Rahman, John See, and Chiung Ching Ho. Action recognition in low quality videos by jointly using shape, motion and texture features. In *Signal and Image Processing Applications (ICSIPA), 2015 IEEE International Conference on*, pages 83–88. IEEE, 2015.

[179] Blake Randolph and Shiffrar Maggie. Perception of human motion. *Annual review of psychology*, 58 :47–73, 2007.

[180] Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction : a survey. *Artificial Intelligence Review*, 43(1) :1–54, 2015.

[181] Daniele Ravı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1) :4–21, 2017.

[182] Brian Reily, Fei Han, Lynne E Parker, and Hao Zhang. Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction. *Autonomous Robots*, 42(6) :1281–1298, 2018.

[183] Zhou Ren, Jingjing Meng, Junsong Yuan, and Zhengyou Zhang. Robust hand gesture recognition with kinect sensor. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 759–760. ACM, 2011.

[184] Alejandro Reyes, Alfonso Alba, and Edgar R Arce-Santana. Optical flow estimation using phase only-correlation. *Procedia Technology*, 7 :103–110, 2013.

[185] Karl Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP-Image Understanding*, 59(1) :94–115, 1994.

[186] Yong Rui, Thomas S Huang, and Shih-Fu Chang. Image retrieval : Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1) :39–62, 1999.

[187] J Russell. Google's alphago ai wins three-match series against the world's best go player. retrieved november 5, 2018, 2017.

[188] M. S. Ryoo. Human activity prediction : Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043, Nov 2011.

[189] D.L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology (2nd Edition)*. Worth, New York, 2011.

[190] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions : a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[191] Makito Seki, Hideto Fujiwara, and Kazuhiko Sumi. A robust background subtraction method for changing background. In *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.*, pages 207–213. IEEE, 2000.

[192] Mubarak Shah and Ramesh Jain. *Motion-based recognition*, volume 9. Springer Science & Business Media, 2013.

[193] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[194] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth

images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12) :2821–2840, 2013.

[195] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[196] Petr Somol, Pavel Pudil, Jana Novovičová, and Pavel Paclık. Adaptive floating search methods in feature selection. *Pattern recognition letters*, 20(11-13) :1157–1163, 1999.

[197] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101 : A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv :1212.0402*, 2012.

[198] V Sugumaran, V Muralidharan, and KI Ramachandran. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical systems and signal processing*, 21(2) :930–942, 2007.

[199] Evan A Suma, David M Krum, Belinda Lange, Sebastian Koenig, Albert Rizzo, and Mark Bolas. Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit. *Computers & Graphics*, 37(3) :193–201, 2013.

[200] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015.

[201] Christian Thurau and Václav Hlaváč. Pose primitive based human action recognition in videos or still images. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[202] Philip HS Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *International workshop on vision algorithms*, pages 278–294. Springer, 1999.

[203] Alexander Toshev and Christian Szegedy. Deeppose : Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[204] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[205] Nikolaus F Troje, Cord Westhoff, and Mikhail Lavrov. Person identification from biological motion : Effects of structural and kinematic cues. *Perception & Psychophysics*, 67(4) :667–675, 2005.

[206] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities : A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11) :1473, 2008.

[207] Pavlo V Tymoshchuk. A discrete-time dynamic k-winners-take-all neural circuit. *Neurocomputing*, 72(13-15) :3191–3202, 2009.

[208] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6) :1510–1517, 2017.

[209] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6) :1510–1517, 2018.

[210] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10) :983–1009, 2013.

[211] Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action recognition by dense trajectories. 2011.

[212] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[213] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition : A survey. *Pattern Recognition Letters*, 119 :3–11, 2019.

[214] Jue Wang, Anoop Cherian, and Fatih Porikli. Ordered pooling of optical flow sequences for action recognition. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 168–176. IEEE, 2017.

[215] Jun Wang. Analysis and design of a k-winners-take-all model with a single state variable and the heaviside step activation function. *IEEE Transactions on Neural Networks*, 21(9) :1496–1506, 2010.

[216] Li Wang, Li Cheng, Tuan Hue Thi, and Jian Zhang. Human action recognition from boosted pose estimation. In *Digital Image Computing : Techniques and Applications (DICTA), 2010 International Conference on*, pages 308–313. IEEE, 2010.

[217] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

[218] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv :1507.02159*, 2015.

[219] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks : Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[220] Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158 :43–53, 2018.

[221] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016.

[222] Ying Wang, Kaiqi Huang, and Tieniu Tan. Human activity recognition based on r transform. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[223] David H Warren and Edward R Strelow. *Electronic spatial sensing for the blind : contributions from perception, rehabilitation, and computer vision*, volume 99. Springer Science & Business Media, 2013.

[224] Daniel Weinland and Edmond Boyer. Action recognition using exemplar-based embedding. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[225] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conference on Computer Vision*, pages 635–648. Springer, 2010.

[226] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2) :224–241, 2011.

[227] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder : Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 780–785, 1997.

[228] Zuxuan Wu, Yu-Gang Jiang, Xi Wang, Hao Ye, Xiangyang Xue, and Jun Wang. Fusing multi-stream deep networks for video classification. *arXiv preprint arXiv :1509.06086*, 2015.

[229] Lu Xia and JK Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841, 2013.

[230] Chew-Yean Yam and Mark Nixon. Gait recognition, model-based. *Encyclopedia of Biometrics*, pages 799–805, 2015.

[231] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[232] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 804–811, 2014.

[233] Angela Yao, Juergen Gall, and Luc Van Gool. A hough transform-based voting framework for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2061–2068. IEEE, 2010.

[234] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 492–497, 2009.

[235] Yang Yi, Yang Cheng, and Chuping Xu. Mining human movement evolution for complex action recognition. *Expert Systems with Applications*, 78 :259–272, 2017.

[236] Sangho Yoon, Chee Sun Won, Kyungsuk Pyun, and Robert M Gray. Image classification using gmm with context information and with a solution of singular covariance problem. In *Data Compression Conference, 2003. Proceedings. DCC 2003*, page 457. IEEE, 2003.

[237] Ting Yu, Cha Zhang, Michael Cohen, Yong Rui, and Ying Wu. Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models. In *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*, pages 1–8. IEEE, 2007.

[238] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets : Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[239] Baochang Zhang, Yun Yang, Chen Chen, Linlin Yang, Jungong Han, and Ling Shao. Action recognition using 3d histograms of texture and a multi-class boosting classifier. *IEEE Transactions on Image processing*, 26(10) :4648–4660, 2017.

[240] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. A key volume mining deep framework for action recognition. In *Computer Vision and*

Pattern Recognition (CVPR), 2016 IEEE Conference on, pages 1991–1999. IEEE, 2016.

[241] Yingying Zhu, Nandita M Nayak, and Amit K Roy-Chowdhury. Context-aware activity recognition and anomaly detection in video. *Selected Topics in Signal Processing, IEEE Journal of*, 7(1) :91–101, 2013.